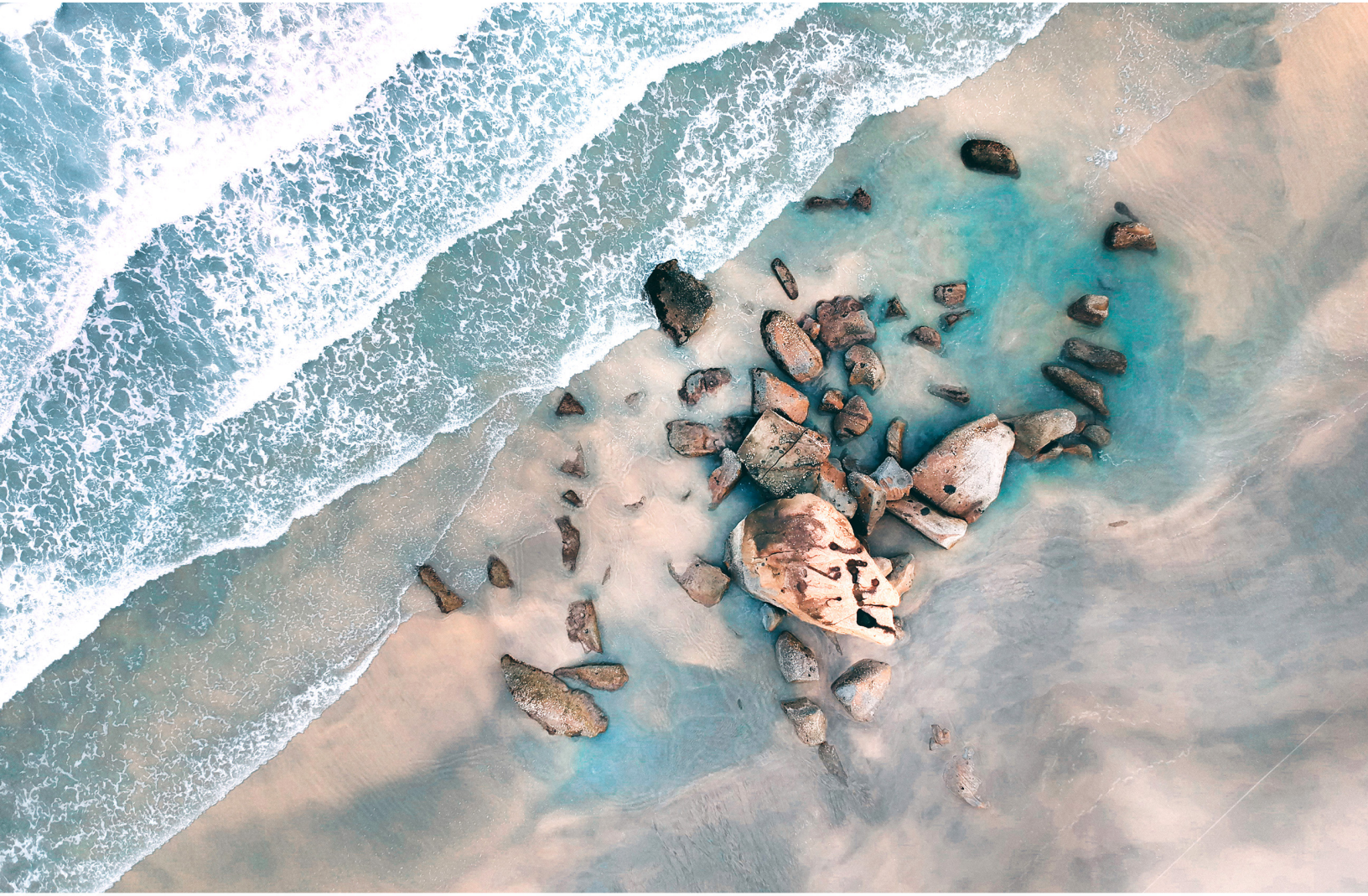


Philippine Information Technology Journal

Vol. XIII No. 1
January - June 2020
ISSN 2012-0761

Published by
COMPUTING SOCIETY OF THE PHILIPPINES
http://www.csp.org.ph

Philippine Information Technology Journal

A Publication by the Computing Society of the Philippines

VOLUME 13 ◦ NUMBER 1 ◦ JANUARY – JUNE 2020 ◦ ISSN 2012-0761

CALL FOR PAPERS

The Philippine Information Technology Journal (PITJ), founded in 2008, is a publication of the Computing Society of the Philippines (CSP), Inc. It appears twice yearly and publishes original peer-reviewed papers on Information Technology, Information Systems, and other applications of Computer Science.

All contributed papers must be written concisely in English and must be submitted electronically to:

The Editors, Philippine Information Technology Journal
Vidal Tan Hall, University of the Philippines Diliman
Quezon City 1101 Philippines

email: subscription@ittc.up.edu.ph

Each paper should have a maximum of fifteen pages, including an abstract of up to 200 words. Submissions may be in Microsoft Word (.doc or .docx) format, LibreOffice Writer (.odt) format, or in LaTeX. Authors submitting in Microsoft Word or LibreOffice Writer format are also required to submit the same paper in PDF. Authors submitting in LaTeX are required to submit all other auxillary files such as graphic files and LaTeX cls and sty files. Authors must typeset their papers using the ACM templates available at http://www.acm.org/sigs/pubs/proceed/template.html.

All references must be listed alphabetically with respect to the first author's last name and cited in the text by square-bracketed numbers. When applicable, each reference item should include the name of the authors, title of the article, journal name, volume, number, month, year, publisher, and pages.

Papers that have been published or submitted to other journals or books will not be considered for publication in the PITJ. Revised or extended versions of papers published in conference, workshop, or symposium proceedings may be submitted to PITJ with proper acknowledgment and cover notes. The PITJ owns the copyright to the technical contributions it published.

SUBSCRIPTION

The annual subscription price is PHP 1,000.00 or USD 40.00 (excluding postage and freight charges). All subscription correspondences should be addressed to: subscription@ittc.up.edu.ph

FOREWORD

The Philippine Information Technology Journal (PITJ) Volume 13 No. 1 showcases selected papers that were presented in the 2020 WORKSHOP ON COMPUTATION: THEORY AND PRACTICE (WCTP 2020). This is the tenth workshop iteration organized by the Tokyo Institute of Technology, The Institute of Scientific and Industrial Research-Osaka University, the University of the Philippines-Diliman, and De La Salle University-Manila. Whose focus is on the theoretical and practical approaches to computation. This workshop aims to present the latest developments by theoreticians and practitioners in academe and industry working to address computational problems that can directly impact the way we live in society.

The WCTP 2020 featured presentations of prominent researchers as well as presentations of research papers selected by members of its Program Committee. They come from highly distinguished institutions in Japan and the Philippines. Their expert knowledge and research experience will certainly provide high-quality reviews from which future submissions can benefit.

Thank you and congratulations to all authors, Adelante!

GREGG VICTOR D. GABISON
Editor in Chief

Philippine Information Technology Journal

A Publication by the Computing Society of the Philippines

VOLUME 13 ◦ NUMBER 1 ◦ JANUARY – JUNE 2020 ◦ ISSN 2012-0761

CONTENTS

Development of an Automatic Speech Recognizer for Filipino-Speaking Children

Matthew Ralph C. Briones, Charlotte M. Cai, Eduardo Emanuel C. Te, Ronald M. Pascual
College of Computer Studies
De La Salle University Manila, Philippines
matthew_briones@dlsu.edu.ph, charlotte_cai@dlsu.edu.ph, eduardo_te@dlsu.edu.ph, ronald.pascual@dlsu.edu.ph

ABSTRACT
In this paper, we present the development of a continuous speech automatic speech recognizer (ASR) for Filipino speaking children using hidden Markov modeling (HMM) approach. A children's read speech corpus in Filipino language (CFSC) was used for training and testing the system. The initial speech corpus containing eight hours of speech data has only one and a half hour of time-aligned phoneme-level transcriptions. In order to transcribe the entire speech corpus, an automatic phoneme-level transcriber based on forced alignment method was created. Both the forced aligner and the continuous speech ASR were developed from phonetic hidden Markov model-based acoustic model. Using bigram language model from the text passages employed in the speech recordings, the ASR system achieved an overall recognition accuracy of 57.9% based on a five-fold cross-validation test. Although the result presented here serves as a baseline for ASR for Filipino speaking children, future directions for improving the result may include phoneme chart revision and the use of hybrid DNN-HMM approaches as more speech data become available.

CCS Concepts
• Computing methodologies→Artificial Intelligence→Natural language processing→ Speech recognition

• Applied Computing→Education→Computer-assisted instruction

Keywords
Automatic speech recognition; Filipino speech; speech technology for children; automatic transcription; forced alignment;

1. INTRODUCTION
Automatic speech recognition (ASR) technologies have been available for quite a while now and are being used throughout the world. For example, Google ASR claims a recognition accuracy of at least 95% in the English language, which can be considered as accurate as a human being when decoding spoken words. ASR systems designed for children however are scarce compared to ASR systems for adult speakers, and their performances are generally not as good as compared to ASR systems for adult. Currently there are no ASR systems that are specifically made for Filipino speaking children. The closest available system to a Filipino ASR was from the study by Pascual and Guevara [1].

Developing an ASR system for children can be challenging because the ways that children articulate vowel and consonant sounds are different as compared to the way adult speakers do. Since children have a limited speech production proficiency in the language they speak, they are more prone to grammatical and pronunciation mistakes. [2][14] There is also high level of variability in children's speech features compared with adults' speech. [14] Another challenge is the lack of available children speech data or speech corpus. [13] One state-of-the-art ASR for Filipino language is the Google ASR. The accuracy of Google ASR has become almost equal to that of a human being, with at least 95% recognition rate for English language. Google ASR's recognition rate for Filipino language however is unspecified. Moreover, the accuracy of Google ASR cannot be claimed as true for the case of children's speech. Thus, if an ASR is to be developed for children's use, the difference between the systems designed for adults and the systems designed for children must also be understood.

Currently the only single existing Filipino children speech corpus available to be used has some issues too. Pascual and Guevara [1] reported that for the children's Filipino speech corpus (CFSC), nearly all the children participants are native Filipino speakers. The researchers classified the speech corpus into two parts: (1) one part containing good reading pronunciations; and (2) the other part containing examples of actual reading miscues and disfluencies. The CFSC also provides data sets, such as: the training data set for the generation of speech models, reference speech features set extracted from good pronunciations, an offline test set for the evaluation of the reading miscue detector (RMD), and a data set for the analysis of actual reading miscues found in Filipino children's speech.

To be able to develop a high accuracy ASR system, the system needs a lot of speech data. This required the authors of this study to transcribe more speech data. Time-aligned phoneme-level transcriptions are required to generate the speech models to be used for the development of the ASR.

For this study, the goal is to create a word-level automatic speech recognizer for continuous children's speech in Filipino. To be able to address the need for more transcribed data for the FCSC, a Viterbi forced aligner using phonetic hidden Markov modeling- or HMM-based acoustic models are utilized. The development and evaluation of the forced aligner will be discussed in the next sections. Moreover, section 2 also discusses in more details the speech corpus and models used in this study. Section 3 presents the results obtained using the different systems and configurations

employed in this study. Finally, section 4 presents our conclusions and recommendations for future work.

2. METHODOLOGY

2.1 Filipino Children Speech corpus

The children speech corpus contains eight hours' worth of children speech data. The content of the speech recordings are text passages from nine Filipino short stories. These speech recordings were used for training and testing the ASR system.

The contents of the CFSC were divided into two parts. The first part has 5 hours' worth of speech data. The aforementioned part is composed of speech collected from 37 students in grades 2 to 5. Twenty students are boys, and seventeen students are girls. The students in the first part of the speech corpus are the students with good reading abilities and good pronunciation skills. The second part of the speech corpus has 3 hours' worth of speech data. The second part is composed of 20 students in grades 1 to 3. Nine of the students are boys and eleven are girls. The students in the second part of the speech corpus are the students that had reading miscues during the recording.

The corpus used for this study contains speech collected from 43 students and has two parts. The first part is the originally transcribed data which is composed of 7 speakers. The second part of the speech corpus contains the newly transcribed data and is composed of 36 speakers. For all sets, each student has their own folder with their recorded speech saved in a '.wav' audio file format sampled at 16 KHz. Each folder contains multiple audio files where each audio file contains one sentence. The folders also include text files containing the text passages read, and other information about the speaker. There are some instances of the same sentences being recorded again as the student had mispronounced or misread some words in the sentence. These recordings with reading errors were also saved in the database along with the corrected version.

The final version of the speech corpus used in this paper contains a total of 3,323 audio files with a total length of around 5.07 hours, which translates to 5 hours, 4 minutes, and 18 seconds. For the originally transcribed data, there were 516 audio files, while for the newly transcribed data, there were 2,807 audio files. Figure 1 and Figure 2 shows the total number of phoneme occurrences in the originally transcribed data and newly transcribed data, respectively.

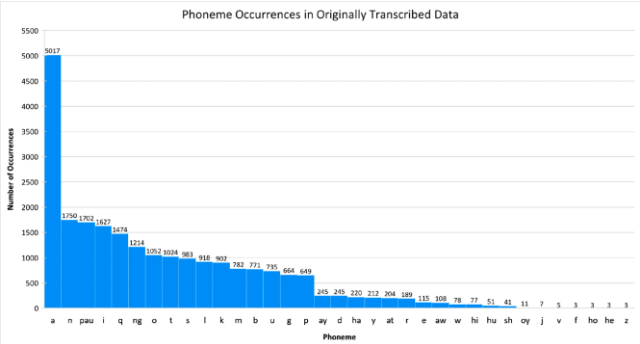


Figure 1. Histogram of phoneme occurrences in the originally transcribed data.

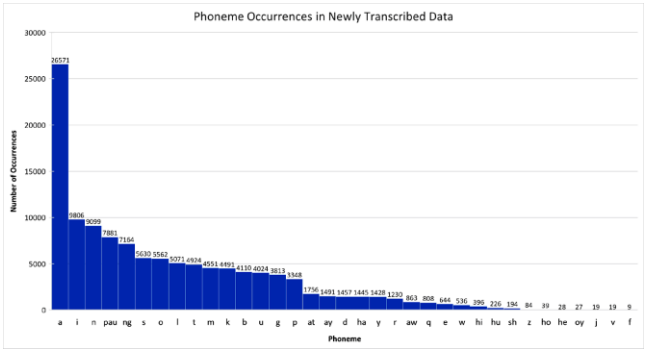


Figure 2. Histogram of phoneme occurrences in the newly transcribed data.

2.2 Speech Transcription

The speech transcription process chosen for this study was the combination of manual and automatic transcription. The automatic transcription was done first using one of the functions of the hidden Markov modeling toolkit (HTK). Then, the researchers manually checked and corrected the alignments produced by the automatic transcriber. It is important to make sure that the transcription procedure is done correctly in order to ensure that the phoneme-level speech models would result to optimum recognition accuracy.

The phonemes used in this study, as well as their classifications are as shown in the phoneme chart in Table 1. This phoneme chart contains thirty-three phonemes and each of these phonemes is based on the words that appear in the text passages. The word-level orthographic transcriptions are based on the text passages, which are the nine Filipino short stories employed during the recording. The nine text passages have a total of 2,169 words, and out of these, 650 words are unique.

Table 1. Phoneme Chart [1]

Phone Class	Phones / Diphones
Stop	/p/, /b/, /t/, /d/, /k/, /g/, /q/ (glottal stop)
Fricative	/f/, /v/, /s/, /z/, /sh/
Affricate	/j/
Nasal	/m/, /n/, /ng/
Lateral Liquid	/l/
Retroflexed Liquid	/r/
Glide	/w/, /y/
Vowels	/a/, /e/, /i/, /o/, /u/
Diphones	/ha/, /he/, /hi/, /ho/, /hu/, /at/, /aw/
Pause/Silence	/pau/

The phoneme chart that is used for this study is the same phoneme chart used in the study of Pascual and Guevara [1], for the simple reason of allowing the authors to use the available time-aligned phoneme-level transcriptions worth one and a half hours. These

available transcriptions were used to bootstrap the automatic transcriber.

2.3 Feature Extraction Module

The chosen method used for the feature extraction module is the Mel-frequency cepstral coefficients (MFCC). The MFCC is a feature commonly used in ASR. The main purpose of the Mel frequency cepstral coefficient is to extract features of an audio signal that allows identification of linguistic content while disregarding unnecessary factors such as unwanted noises and other unnecessary information such as voice quality, gender of a speaker, emotion, etc. MFCC is also said to accurately represent the shape of the vocal of the short time power spectrum envelope. MFCC is known for its accuracy in estimating the speech parameters and efficient computational model of speech. [5]

The number of coefficients used in this study is 39. Choosing the number of coefficients is dependent on the required level of accuracy, however the number of coefficients typically used is 39. [5]

The feature extraction module takes the raw speech recordings from the speech corpus as input. A speech recording contains a recording of a sentence from a short story. Next, a configuration file that contains all the conversion parameters is created. The number of coefficients is set in the configuration file to be 39, that is, 13 MFCC zero mean static coefficients (Z) including energy (E), plus 13 delta coefficients (D), plus 13 acceleration coefficients (A). Based on our experiments, the aforementioned configuration resulted in the highest accuracy. Once the configuration parameters are set, the MFCC files are then created.

The TARGETRATE, which is the frame period, is set to 10 milliseconds. The MFCCs were saved in a compressed format with an added CRC checksum. The windowing method employed the Hamming window and a first-order pre-emphasis was applied using a coefficient of 0.97.

2.4 Acoustic Modeling

The chosen model for this study is the hidden Markov model (HMM). A Markov model is a model that depicts every observable event as a state. Speech signals are normally continuous and are difficult to determine when and how an abstract speech code transitions to another. Therefore, an explicit, definitive observation of a state sequence cannot be assumed; each state cannot be associated with an observable event. To add flexibility to the model, the outcomes or observations of the model are assumed to be a probabilistic function of each state. The state sequence is hidden or is not directly observable; it can only be approximated from the sequence of observations produced by the system, hence the term hidden Markov model [6]. The system uses a four-state left-right HMM. Based on our initial experiments, a four-state HMM resulted to relatively optimum phoneme recognition rate for our phoneme set without requiring much computation. Unlike the more common three-state HMMs, a four-state seem to give better results due the presence of many diphones in the phoneme set used in this study. Figure 3 shows the state diagram of a four-state left-right HMM.

Table 2 shows the phoneme-level recognition accuracy of each N-state HMM. Recognition accuracy is calculated by subtracting the number of insertions from the correctly recognized phonemes, then dividing the result by the total number of phonemes.

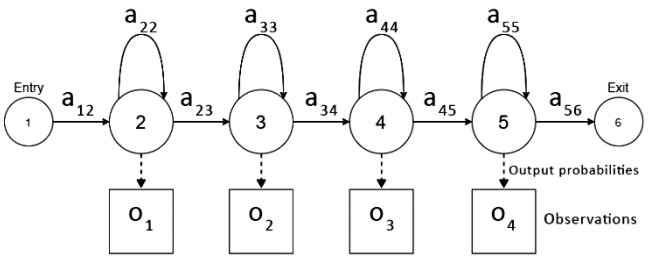


Figure 3. A 4-state left-right HMM.

The 4- and 5-state HMMs have fairly close accuracies, but four-state is still preferred as its computational cost is lesser. More than five states can also be used, but there are no studies or theories that prove that having more than five states will further significantly improve the accuracy of such ASRs.

Table 2. Phoneme recognition accuracy breakdown for each N-state HMM.

	3-State	4-State	5-State
Correctly Recognized	77539	77446	76263
Deletions	15832	20727	23661
Substitutions	51087	46285	44534
Insertions	23660	15365	12600
Total Occurrences	144458	144458	144458
Accuracy	37.30%	42.98%	44.07%

The acoustic modeling module is responsible for training the system by first creating a flat start prototype model that initializes the parameter of the HMM, that is, by setting all the Gaussians in a given HMM to have the same mean and variance. After setting an initial HMM, the next step is to take the MFCC files from the feature extraction module, the time-aligned phoneme transcriptions and the HMM list. The Baum-Welch algorithm is the re-estimation technique used for training the acoustic models. The Baum-Welch algorithm trains the HMMs by maximizing the probability that a certain state sequence is observed given the acoustic evidence or observations from the training data or examples.

2.5 Automatic Transcriber

The automatic transcriber system for this study uses the method called Viterbi forced alignment. Forced alignment is the process of finding, for each fragment of the data transcript, the time interval of the speech file containing the spoken text of the fragment. Having forced alignment tools can help minimize the time needed to align transcriptions compared to aligning them manually. Forced alignment is also better at achieving results with consistency and replicability [7]. The main algorithm for forced alignment, which is the Viterbi algorithm, minimizes error probability by comparing likelihoods of a set of state transitions that can occur, and then deciding which transition has the highest

probability of occurrence. Thus, the path that has the highest probability would be chosen by the algorithm. In forced alignment, the Viterbi algorithm will find the most likely alignment or sequence of the generated Hidden Markov Models to the speech data.

Using the feature vectors and reference phoneme sequence representing the words in the speech recording, the Viterbi algorithm is then used to time-align the speech signal and the reference phonemes. The result is a label file that contains the time-aligned speech transcriptions that indicates the time when each sound starts and ends.

2.6 Phoneme Recognizer

The phoneme Recognizer was created using the HMM acoustic models that were trained using the CFSC data. The phone-based HMM models were estimated eight times. The accuracy of the system started to plateau at re-estimation 5 (HMM5) up to re-estimation 8 (HMM8). This can be seen from Figure 4.

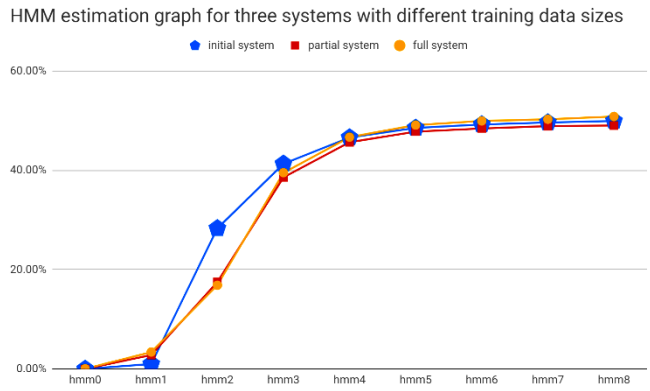


Figure 4. HMM estimation graph for three systems with different training data sizes.

Since there are no existing ASRs for Filipino speaking children and only a limited amount of time-aligned phoneme-level transcriptions are available, the authors decided to first develop a phoneme-level recognizer system that iteratively will have increasing training data size as more transcriptions become available. The three systems considered were called: the initial system, the intermediate/partial system and the final/full system. The difference between each system was the size of speech data used.

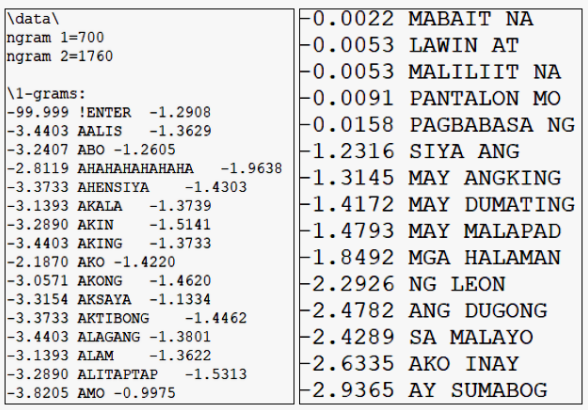
For all the three systems, the speech data set used was organized according to speaker. Thus, for each speaker involved in each system, all the audio files with his/her speech would be used. For the initial system, the speech data used for training and testing were the originally transcribed data provided by the CFSC. The intermediate system used half of the non-transcribed data from the CFSC combined with the originally transcribed data. As for the final system, the speech data used for training and testing was the entire CFSC.

2.7 Language Model

The language model used for the word recognizer was a bigram language model. A bigram language model is used for approximating the probability of a word based on the previous word. The bigram language model can help the system narrow down the list of words from a word dictionary by predicting the next word to appear based on the previous word that was recognized.

The left side of Figure 5 shows a snippet of the 1-gram section of the bigram file used for the system, while the right side shows a portion of the 2-gram section. For the 1-gram section, n-gram 1 shows how many words are in the 1-gram, while n-gram 2 shows how many pair words are in the 2-gram. The first column for both 1-gram and 2-gram in Figure 5 represents the probability of the word in base-10 logs (the closer the value to 0, the higher the probability for the word or the word pair). The third column in the 1-gram shows the back-off weight, which is used when determining which n-gram to use in case a back-off occurs. A back-off occurs whenever an utterance was not included in an n-gram model.

To further explain the examples in Figure 5, consider the15-word pairs in the right side of the figure. The first five word pairs are considered as the most common word pairs found in the speech data. For example, the pair ‘MABAIT NA’ appeared 10 times in story 8, hence the reason why the bigram model considered the aforementioned word pair as one of the most common instances. The next five-word pairs may be considered to have a relatively moderate frequency or occurrences in the speech data. The last five-word pairs are considered the word pairs with the least number of examples found from the speech data. The word pair ‘AY SUMABOG’ appeared only once in the whole speech corpus, hence the reason why the bigram model assigned it with the lowest probability.



```
\data\
ngram 1=700
ngram 2=1760

\1-grams:
-99.999 !ENTER  -1.2908
-3.4403 AALIS   -1.3629
-3.2407 ABO -1.2605
-2.8119 AHAHAHAHAHAHA   -1.9638
-3.3733 AHENSIYA    -1.4303
-3.1393 AKALA   -1.3739
-3.2890 AKIN    -1.5141
-3.4403 AKING   -1.3733
-2.1870 AKO -1.4220
-3.0571 AKONG   -1.4620
-3.3154 AKSAYA  -1.1334
-3.3733 AKTIBONG    -1.4462
-3.4403 ALAGANG -1.3801
-3.1393 ALAM    -1.3622
-3.2890 ALITAPTAP   -1.5313
-3.8205 AMO -0.9975
```

```
-0.0022 MABAIT NA
-0.0053 LAWIN AT
-0.0053 MALILIIT NA
-0.0091 PANTALON MO
-0.0158 PAGBABASA NG
-1.2316 SIYA ANG
-1.3145 MAY ANGKING
-1.4172 MAY DUMATING
-1.4793 MAY MALAPAD
-1.8492 MGA HALAMAN
-2.2926 NG LEON
-2.4782 ANG DUGONG
-2.4289 SA MALAYO
-2.6335 AKO INAY
-2.9365 AY SUMABOG
```

Figure 5. Selected portion of the bigram language model.

2.8 Accuracy Test for the Forced Aligner

To measure the accuracy of the alignments generated by the forced aligner, the absolute difference of the start and end time boundaries for each of the phonemes from both the automatically generated alignments and the manually generated (or hand-corrected) alignments were calculated. Finding the absolute difference of the time boundaries of each phoneme can help point out if the time boundaries of the generated alignments are within a certain time threshold of the time boundaries of the manually fixed alignments.

The goal is to use the time difference to see if the time boundaries of the generated alignments can fit into a certain time threshold of the time boundaries of the manually corrected alignments. [8][9] The time tolerances or time threshold values chosen for this study are the following: 10msec, 20msec, 30msec,

and 40msec. [8][9] For each time threshold values, the time difference of the boundaries can be part of the non-error rate or the error rate. A time difference is considered as part of the non-error rate if the time difference is within the threshold time. If the time difference exceeds the time threshold then it will be considered part of the error rate.

2.9 Accuracy Test for Phoneme and Word Recognizer

The three systems were tested on both phoneme-level and word-level. A K-fold cross-validation test was conducted for phoneme-level, with 5 folds for all systems except for the initial system which reached up to 6 folds. For all systems, each test data contained isolated data where that data was not part of the training data for generating the HMMs. The initial system consists of 6 speakers and each speaker was tested with a K-fold of up to 6. The intermediate system consists of 26 speakers and five separate tests with 5-fold cross-validation were conducted. Each test was composed of a group of speakers. Lastly, the final system consists of all 43 speakers and five separate tests with 5-fold cross-validation were conducted.

Similarly, the word-level ASR systems were also all tested with 5 separate test groups. Each test group's recognition accuracies were then averaged. The average word error rates (WER) were also calculated.

3. RESULTS AND DISCUSSION

3.1 K-fold Test

The result of the first K-fold tests involves data from using the initial system consisting of 440 WAV files containing speech data from six speakers. These speakers read 7 stories out of the 9 available stories in the FCSC.

Looking at the results in Table 3, the highest recognition accuracy from the whole initial system can be seen in K3 with the test data as speaker 11 with 60.25% phoneme recognition accuracy. It achieved this recognition accuracy because the speech data spoken by speaker 11 has more common instances with most of the training data. It is also due to speaker 11 being part of the training data of the 3rd fold, making it easier for the initial system to recognize the phonemes uttered by speaker 11.

In each k-fold test, there are these "isolated test data", wherein these data are not part of the training data of the HMM model. The isolated test data are those percentage values colored in red shown in Table 3. For these isolated test data, the highest recognition accuracy of 44.72% can be found in k3 where the isolated test data is from speaker 7. The reason why isolating speaker 7 as a test data produced a high accuracy is that it had more common instances with the majority of the training data.

The next k-fold test was using the Intermediate system. The data available in the intermediate system consists of 26 unique speakers with a total of 5500 WAV files. The data set consist of the following: speaker numbers 21 up to 39 of the newly transcribed data, and speakers 3, 5, 7, 9, 11, 18, and 19 of the originally transcribed speakers. These speakers have read 8 stories out of the 9 available stories in the FCSC.

Table 3. Results of K-Fold Accuracy Test for The Initial System (phoneme-level).

K	Speaker 3	Speaker 5	Speaker 7	Speaker 11	Speaker 18	Speaker 19
1	**44.06%**	46.48%	53.76%	57.96%	51.90%	42.60%
2	47.32%	**38.7%**	53.31%	54.59%	52.35%	39.34%
3	44.35%	49.80%	**44.72%**	60.25%	49.11%	43.41%
4	44.24%	38.51%	51.81%	**38.52%**	50.43%	36.17%
5	44.17%	51.40%	50.07%	59.05%	**43.73%**	44.47%
6	46.99%	47.09%	53.67%	57.26%	52.22%	**38.28%**

As shown in Table 4, the overall highest recognition accuracy of 54.37% for the intermediate system is in the k4 test with T1 as the test data. This can be explained by the fact that the speakers in the test group T1 primarily consist of the same stories and instances as the training data. T1 speakers mostly read stories 1, 2, 4, and 5 which is 77% of the training data. That is why even when T1 was considered as the isolated data in the k1 test, it still garnered a 50.61% accuracy, which is the highest recognition accuracy amongst the other isolated data tests.

Table 4. Results of K-Fold Accuracy Test for The Intermediate System (phoneme-level).

K	T1	T2	T3	T4	T5
1	**50.61**	52.21	47.46	49.33	41.68
2	54.04	**46.48**	47.91	47.84	41.78
3	53.27	51.73	**45.37**	48.06	41.77
4	54.37	52.95	48.66	**48.0**	42.53
5	53.60	52.39	47.46	48.62	**38.23**

The last system that went through the k-fold test was the final system. The data available in the final system consists of 43 speakers and a total of 9,942 WAV files. These speakers for this data set are the following: speakers numbers 21 up to 59 (except speakers 43, 45, and 53) of the newly transcribed data, and speakers 3, 5, 7, 9, 11, 18, and 19 of the originally transcribed data. All 9 available stories in the FCSC were read by the aforementioned speakers.

As shown in Table 5, the overall highest recognition accuracy of 55.79% was obtained for the full system in the k3 test with T4 as the test data. Similar to previous systems, the speech data in group T4 had a lot of similar instances with the training data. For instance, all the speakers in T4 read story 1 which is the most common story for the whole speech data, making the phonemes of the story easier for the system to recognize since the system has a lot of instances to refer to. The highest recognition accuracy of 53.24% was obtained in the k2 test with T2 as the isolated test data. Similar to the T4 group, T2 was also a group that had a lot of similar instances with the training data. That is why even though T2 was part of the isolated test data, the system still achieved a high recognition rate.

Table 5. Results of K-Fold Accuracy Test for The Final System (phoneme-level).

K	T1	T2	T3	T4	T5
1	**32.56**	53.07	52.97	55.02	48.68
2	42.23	**53.24**	52.73	55.13	50.63
3	44.25	54.99	**51.5**	55.79	52.04
4	42.92	53.03	52.54	**52.4**	48.85
5	42.17	54.2	53.56	54.55	**47.51**

3.2 Phoneme Recognizer

The final system in the phoneme-level tests achieved an overall average recognition accuracy of 50.26%, which is the highest among all the three systems namely the initial, the intermediate, and the final system. The summary of results for phoneme recognition tests are shown in Table 6. The result in Table 6 implies that having more training data increases the accuracy of the system. Note that the final system had the highest number of training data among all systems. The final system had 43 speakers as training data, while the intermediate system was trained with 26 speakers, and the initial system only had 6 speakers as training data. It can be noted however in Table 6 that the increase in recognition rate from initial to intermediate system, as well as from intermediate to final system, are not significantly high.

Table 6. Average phoneme recognition accuracy for the three systems used in the study

System	Average Recognition Accuracy
Initial System	47.84%
Intermediate System	48.25%
Final System	50.26%

3.3 Forced Aligner

As mentioned in the previous sections, a Viterbi forced aligner was created for the purpose of automatically transcribing the remaining non-transcribed speech data in the entire speech corpus. The resulting alignments from the system seemed mostly accurate when viewed in Praat. Figure 5 shows an example of an audio file with a good alignment. However, as mentioned, the alignments were not perfect, and Figure 6 shows one of the resulting alignments that had heavy misalignments. In Figure 6, the force aligner recognizes the second phoneme 'p' as part of the pause or silent 'pau' phone, while the time frame of 'p' is actually small and it occurs only for a short time before the third phoneme 'a'. Commonly encountered problems, usually found at the beginning of the audio recordings, includes background noises that were picked up during the recording session, unexpected noise such as breathing sound, coughing, and noise when clicking the start recording button.

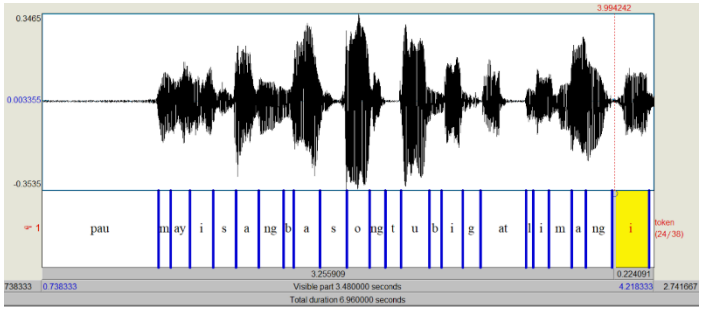


Figure 5. Good audio with no interfering sound. No misalignment, no sliding

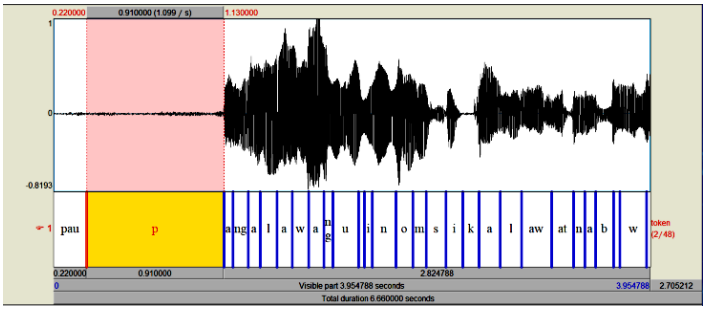


Figure 6. Sample audio that had heavy misalignment starting at highlighted phoneme /p/.

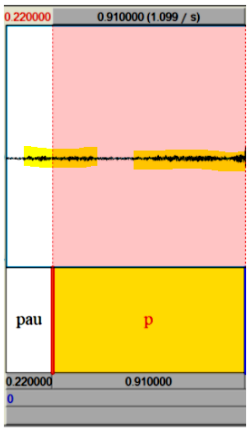


Figure 7. Sample audio that had heavy misalignment starting at highlighted phoneme /p/".

Figure 7 shows an example of a problem area encountered during the alignment process; the problem area is highlighted in yellow. These problem areas can cause the system to have wrong alignment or cause phoneme sliding. An example of this is seen in Figure 6. For both Figures 6 and 7, the first tier (topmost tier) shows the speech audio waveforms, while the second tier shows the phoneme transcription produced by the force aligner.

Based on the results shown in Table 7, at the time difference of 10 milliseconds, the system had 60.50% error. It means that most of the alignments had more than 10 msec time difference compared to the original alignments. However, the percent errors for 20- and 30-millisecond threshold values are down to 41.80% and 30.14% respectively. This shows that the time alignments produced by automatic transcription are relatively not that far off from the alignments produced manually by linguistic experts.

Table 7. Time-alignment error and accuracy rates obtained by the phoneme-level forced aligner across various time difference threshold values.

Time Difference Threshold	%Error	%Accuracy
10msec	63.50%	36.50%
20msec	41.80%	58.20%
30msec	30.14%	69.86%
40msec	26.54%	73.46%

3.4 Word-level Continuous Speech ASR

The final system achieved the highest word recognition accuracy among all three systems. Table 8 shows the average recognition accuracy and word error rates (WER) obtained for each of the three systems namely the initial, the intermediate, and the final system. This can be interpreted in the same way as the results from the phoneme-level tests. That is, more training data generally increases the recognition accuracy of the system. However, it should be noted that the recognition accuracy increases from the intermediate to the final system was much smaller compared to the increase from the initial system to the intermediate system. This may suggest that the system will no longer improve much further by adding more speech data to the final system, and that the current data set in use is sufficient for the system to reach its optimal level.

Comparing the phoneme-level and word-level ASR, it is seen that the word-level ASR achieved a higher overall recognition accuracy of 57.90% as compared to the phoneme-level ASR which only reached an overall recognition rate of 50.26% with the final system. This may be explained by the fact that the word-level ASR used a bigram language mode that helped boost the system's ability to recognize the words uttered, compared to the phoneme-level ASR.

Table 8. Average recognition accuracies and word error rates for each of the three systems employed in the study

System	Average word recognition accuracy	Average word error rate
Initial System	43.91%	56.09%
Intermediate System	56.00%	44.00%
Final System	57.90%	42.10%

To further analyze the results of the word-level recognition, the resulting phoneme and word recognition from a speech audio input was combined and viewed in the application PRAAT. Figure 8 shows the PRAAT view of both results together with the audio input. For this sample the input audio file contains utterances of the words 'ISANG', 'MAYANG', and 'UHAW'. Looking at the phoneme tier, which is the 2nd tier in Figure 8, the phoneme recognition for the words 'MAYANG' and 'UHAW' had some misrecognized phones which were the highlighted 'ha', 'y', and 'w'. The misrecognized 'y' caused the word recognition to not recognize the word 'MAYANG'. Moreover, there was a misrecognized 'ha' which also contributed to the failure of the word recognizer since there is no phoneme sequence 'm ha ay y ng' in the word pronunciation dictionary. However, the reason why the word 'MAYA' was recognized was because of the bigram language model used, since in the bigram language model there is a word pair 'MAYA NG', so the 'ng' between 'y' and 'u' helped the word recognizer to recognize the unknown sequence as 'MAYA'. As for the word 'UHAW', the word was misrecognized because of the insertion of the phone 'w' which could have led the ASR to assume that the word was 'NAUHAW' instead of 'UHAW'. Note that the sequence 'u-w-h-a-aw' is not part of the word pronunciation dictionary, and that the word recognizer had to rely on the language model to recognize what the word was. In the language model, the word 'NAUHAW' had a higher back-off log probability compared to the word 'UHAW'. Since the sequence was only for a certain word, the language model had to back-off to the 1-gram model. Since 'NAUHAW' had a close sequence to the recognized sequence, and at the same time has a higher back-off log probability, 'NAUHAW' instead of the word 'UHAW' was chosen by the ASR.

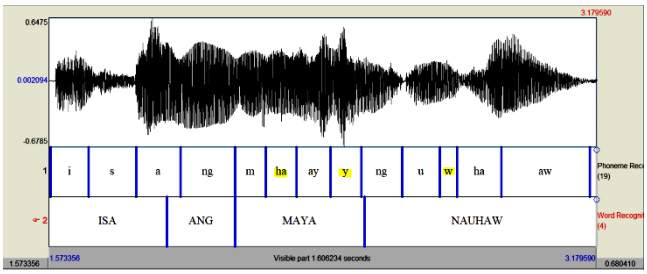


Figure 8 An example of a phoneme recognition and word recognition result.

4. CONCLUSION

The results of all three systems, namely the initial, the intermediate, and the final system, generally had little to no significant improvements in the phoneme recognition accuracies even though the intermediate and final system had more training data. This means that adding more speech data for training the speech models may not help the system to achieve significantly higher recognition accuracies.

For the word-level recognizer, it was seen that there was a significant improvement of about 12% in the recognition rate from the initial to the intermediate system. However, from the intermediate to the final system, the improvement in the recognition rate is only a little less than 2%. This would again suggest that adding more speech data may not help to further significantly increase the recognition accuracy.

Even though the final system has an overall recognition rate that may be considered inferior to those from other studies such as those of Gray et al. [10] and Bautista and Kim [11], there are not enough bases for direct comparison because developing ASR for children is considered to be generally more difficult than developing ASR for adult. [2][14] There is currently no available Filipino ASR for children to compare the study with. Thus, the results from this study can serve as a baseline for any similar studies in the future.

The results of the forced aligner were generally not in huge disagreement with transcriptions manually generated by experts. However, in our experience, noises such as voices of other people

in the background, and the clicking of a button that starts the recording, can significantly affect the forced alignment process causing the resulting alignment to have serious time sliding.

For future work, researchers should explore other modeling approaches such as DNN or hybrid DNN-HMM. It is important to note however that DNN or hybrid DNN-HMM acoustic modeling requires a lot of speech data [12]. In order to make up for the small amount of data in the study in [12], the researchers produced a large population of pseudo-samples from the small speech data.

Another recommendation is for future researchers to consider reviewing the phoneme chart and re-transcribing the speech data if needed. According to our observation, the difference in the transcription convention of the people who generated the initially available transcriptions, versus the transcription style of people who generated the newly transcribed data, had a big effect on the accuracy of the final system. One example of transcriber disagreement is the case of the glottal stop /q/, which happens to have the lowest recognition rate according to our confusion matrix. Modifying the phoneme chart and making it more suitable to the CFSC data may also help in improving the recognition accuracy of the ASR, since we have observed cases where two phoneme sounds such as /it/, /iw/, and /na/ were blending together and are difficult to transcribe separately.

5. REFERENCES

[1] Pascual, R. and Guevara, R. C. 2012. *Developing a children's Filipino speech corpus for application in automatic detection of reading miscues and disfluencies*. TENCON 2012 IEEE Region 10 Conference, Cebu City, pp. 1-6.

[2] Tong, R., Wang, L., and Ma, B. 2017. *Transfer learning for children's speech recognition.* International Conference on Asian Language Processing, December 2017. DOI= 10.1109/IALP.2017.8300540

[3] Prasad, B. and Prasanna, S. 2008. *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2008, pp. 23-47.

[4] Basu, J., Bepari, M., Nandi, S., Khan, S. and Roy, R. 2013. *SATT: Semi-automatic transcription tool*. 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Gurgaon, 2013, pp. 1-6. DOI= 10.1109/ICSDA.2013.6709903

[5] Majeed, S., Husain, H., Samad, S., Idbeaa, T. 2015. *Mel Frequency Cepstral Coefficients (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition: A Comparison Study*." Journal of Theoretical & Applied Information Technology 79.1 (2015).

[6] Jeebun, S., Ramjug-Ballgobin, R., and Al-Ani, T.2015. *Optimal Number of States in Hidden Markov Models and its Application to the Detection of Human Movement*. University of Mauritius Research Journal, 2015. [Online]. Available: https://www.semanticscholar.org/paper/Optimal-Number-of-States-in-Hidden-Markov-Models-to-Jeebun-Ramjug-Ballgobin/680abe2ffbaae0748f09a3e6078e28c379db8781

[7] Malaay, E., Simora, M., Cabatic, R. J., Oco, N., andRoxas, R. E. 2017. *Development of a multilingual isolated digits speech corpus*. Proc. 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA) 2017.

[8] Hosom, J. 2009. *Speaker-independent phoneme alignment using transition-dependent states*. Speech Communication, 51(4), 352–368. doi:10.1016/j.specom.2008.11.003

[9] Stolcke, A., Ryant, N., Mitra, V., Yuan, J., Wang, W., and Liberman, M. 2014. *Highly accurate phonetic segmentation using boundary correction models and system fusion*. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI= 10.1109/icassp.2014.6854665.

[10] Gray, S., Willett, D., Lu, J., Pinto, J., Maergner, P., andBodenstab, N. 2014. *Child Automatic Speech Recognition for US English: Child Interaction with Living-Room-Electronic-Devices*. Proc. 4th Workshop on Child Computer Interaction (WOCCI 2014).

[11] Bautista, J., and Kim, Y. 2014. *An Automatic Speech Recognition for the Filipino Language using the HTK System*. Proc. Int'l Conf. Artificial Intelligence (ICAI'14).

[12] Ghaffarzadegan, S., Bořil, H., and Hansen, J.2017. *Deep neural network training for whispered speech recognition using small databases and generative model sampling*. Int. J. Speech Technology 20, 1063–1075 (2017). https://doi.org/10.1007/s10772-017-9461-x

[13] Rahman, F., Mohamed, N., Mustafa, M., and Salim, S. 2014. *Automatic speech recognition system for Malay speaking children.* 2014 Third ICT International Student Project Conference (ICT-ISPC), IEEE, pp. 79-82.

[14] Russel, M. 2010.*Speech technologies for children*. IEEE Signal Processing Society - STLC Newsletter, New Orleans.

Implementation of Various Machine Learning Techniques to Identify or Characterize Basketball Shots Using OpenPose

Michiko Go
De La Salle University
Manila, Philippines
michiko_go@dlsu.edu.ph

Brent Aldwinne Lim
De La Salle University
Manila, Philippines
brent_lim@dlsu.edu.ph

Charles Navarro
De La Salle University
Manila, Philippines
charles_navarro@dlsu.edu.ph

Enoch Puno
De La Salle University
Manila, Philippines
enoch_puno@dlsu.edu.ph

Fritz Kevin Flores
De La Salle University
Manila, Philippines
fritz.flores@dlsu.edu.ph

***Abstract*—Current technologies are being implemented in various fields in sports science such as basketball. This study analyzes different joints, lengths and angles gathered in basketball shooting motion for shot prediction. The goal of the study is to observe if using machine learning for shot prediction is feasible. To gather the data needed, a program using OpenPose libraries will be used for collecting the skeletal data which are the joints' coordinates in every frame of the shooting motion video. The study would cover two different shot types: free-throw, and three pointer. The participants will be divided into professionals and non-professionals to prepare the gathered data for machine learning, the gathered data will be used in two ways: as an overlayed image of the shooting motion consisting of skeletal points and the gathered data transformed as time series data. To achieve this, the data would undergo preprocessing and time series data transformation techniques to properly prepare the data. Afterwards, different machine learning algorithms would be used to train and test the processed data. Performance metrics are prepared to evaluate each machine learning algorithm used to observe which among the machine learning models yielded the best results.**

***Index Terms*—Machine Learning, OpenPose, Time Series, Basketball, Skeletal Data**

I. Introduction

Sports are a very important part of everyday life. [1] stated that sports "has transformed to the state of a big industry with its advertisement incomes, sponsorships, live broadcasts, club products devoted to the fans, club's stock exchange securities, sport materials sold in the stores, match ticket fees paid by the audiences setting their hearts on the club, combines cards and even with the 'special credit cards of the banks for the fans' and it has come off from its amateur spirit and transformed to a professionalized phenomenon". With the emergence of sports in the lives of people, many conduct studies on sports and focus on how sports affect people's conditions. Sports sciences allow improvement on the performance and health of people on a certain sport. The application of technology, sports science and analytics in sports can be helpful and beneficial to players and newcomers alike. The fusion of sports science and analytics has not only revolutionized the way professional basketball is played, coached, and managed; it also has the potential to revolutionize the way we educate the next generation [2]. The general objective of the study is to implement various machine learning techniques to determine patterns of basketball shots through the use of shot and motion data obtained using a visual sensor. The study's specific objectives are to gather shot and motion data from different basketball shooting forms using OpenPose, to perform feature extraction and data processing from the collected data, and to apply various machine learning algorithms to determine successful and unsuccessful basketball shots.

II. Review of Related Literature

Throughout the years, studies were conducted that are related to basketball and sports through action and shot pattern analysis. These studies primarily focus on wearable sensors, cameras, machine learning, shot and action patterns, and pose recognition.

A study by [3] presented using a wireless hybrid sensor WAA010 that compromises a 3-D Acceleration and Gyroscopic Sensor that analyzes the shooting motion of the athlete. This is used using the waveform of the angular velocity of the athlete. Studies that use sensors for analyzing motion in specific body parts are made in [4] and [5]. [4] used a wrist-worn sensor to analyze a basketball motion. This is used by getting the acceleration of the hand and labeling it as a specific basketball motion. As for the study of [5], the study used a body-worn sensor, more specifically a sleeve, to analyze the free throw shooting form. The sensor was used to transmit data to the hub which included two ZigBee modules and a Micro-SD card. The data transmitted was the time of the shot of each athlete.

[6] uses an OpenPose camera to predict the free throw shot of an athlete. Using the skeletal data of the athlete with the logistic regression, SVM and Xgboost to be used in predicting the shot of the athlete. The OpenPose camera is a device that does not need to be strapped to the player. The athlete can practice freely without any hesitations or disturbance around. This device is similar to a Kinect, the only difference is that the Kinect uses a 3D camera while OpenPose uses a webcam. However, OpenPose can be costly in computing. OpenPose's runtime is proportional to the number of people in the image, making whole-body OpenPose prohibitively costly for multi-person and real-time applications.

The study by [7] focuses on extracting actions from a video recorded using a digital camera and recognizing the action using a dataset containing proposed actions. The preprocessing used is extracting the chromatic frame using a threshold, using the illumination level of the frame, noise and other impairments dilation and erosion is performed, the background is also removed from each frame. Moments extraction is used to categorize the shape and features. The scale, location and rotation invariant moments are used to extract features regardless of size, position and rotation. Raw moments are calculated along the origin of the image, this provides information about properties like area and size of the image. Central moments are invariant to translation of objects in an image, they are computed along the centroid rather from the origin. Scale Invariant Moments are raw moments and central moments depend on the size of object, creating a problem when the same object is compared by both the images are captured from different distances. Rotational invariant Moments are invariant to change in scale and rotation. The study shows that the overall accuracy is 80.8%. The system would be able to recognize medial frames rather than initial or terminal ones. The accuracy can increase to 95% if the medial frames are considered.

A study by [8] aims to use a 3D sensor to build a gesture recognition system for human robot interaction. The approach consists of two layers which are: detection, and tracking. Detection layer is for the extraction of the features, while the tracking layer is for performing temporal data association. These layers were implemented with the help of the Kinect sensor. The steps used in this study are: Feature Extraction and Data Preprocessing.

With all the studies discussed, most of the studies use invasive methods such as invasive and high cost sensors for motion detection and data gathering. Wearable sensors may cause discomfort when worn during motion and movement, thus performance may be affected. High cost sensors can be very costly for the researchers to use for study.

This can be a research opportunity to propose a non invasive and low-cost way of gathering kinesthetic data to analyze shot patterns. The study will use an OpenPose real-time keypoint detection library for body, face, hands, and foot estimation for data extraction from videos of shooting motion.

III. Methodology

A. Overview

Fig. 1. Data Collection Framework

Fig. 2. Feature Extraction Framework

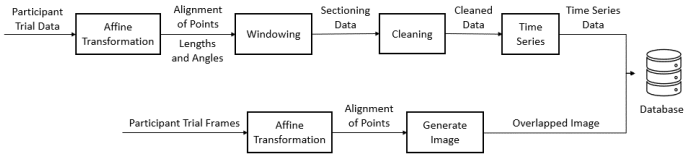


Fig. 3. Preprocessing Framework

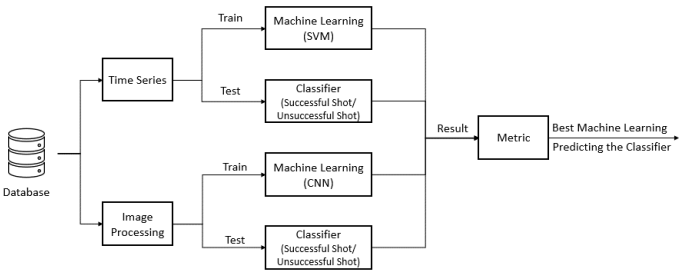


Fig. 4. Machine Learning Framework

To achieve the objectives of the study, the methodology of the study is divided into four phases: (1) data collection, (2) feature extraction (3) preprocessing and (4) machine learning. The Data Collection phase will consist of gathering footage of participants' shooting motions of free throw shots and three point shots. The OpenPose Point Extraction phase will consist of the extraction of frames from the collected videos of the participants and the collection of skeletal data from the videos of participants performing the shooting motion through the OpenPose tool. The preprocessing phase will consist of data cleaning and image processing. Lastly, the machine learning phase will consist of the interpretations of the two preprocessed outputs after feeding them to the machine learning models and metrics to be used to evaluate the performances of the models, to see which one performs best.

B. Data Collection

In data collection, the proponents gathered participant data and online data. The participants involved in the study primarily consist of basketball hobbyists. The online participants for the study consist of video footage of professional basketball players from the NBAand PBA making shot attempts that are found on sites such as YouTube, Facebook and Twitter. The participants in the study are limited to only males, ageing around 18 to 24 years old, and righthanded shooters since right-handed shooters since this is more common among players. The total number of total participants and shots can be seen in Table I and Table II.

	Non-Professional		Professional	
	Free Throw	Three Point	Free Throw	Three Point
Failed Shots	34	30	2	27
Success Shots	46	30	16	83
Total	80	60	18	110

TABLE I
TOTAL SHOTS GIVEN BY PARTICIPANTS

	Free Throw Shot	Three Point Shot
Non-Professional	8	6
Professional	3	11
Total	11	17

TABLE II
TOTAL PARTICIPANTS IN THE STUDY

Players are fully rested to avoid external factors such as fatigue in affecting the result of the shots. Before asking for the participants to perform the shots needed for the study, the proponents prepared consent forms to inform the participants of what information and data the study will be gathering from them. This is to satisfy the requirements from the DLSU Research Ethics Office, for the team to be aligned with research ethical standards involved with the study. upon agreeing upon the consent forms, the participants are asked for the following information: age, height, weight and wingspan. Each participant is asked to perform the following shot types: (1) free throw shot and (2) three point jumpshot. Each participant is placed in front of the camera at a distance of 10-15 feet for the camera to capture enough of the shooting motion of the participant. Each participant is asked to perform 10 trials of the free throw shot and another 10 trials of the three point jumpshot. For the professionals that serve as online participants, the shot type they are making is selected, not both because the footage is only found online. The number of attempts the professionals are making is not fixed to 10 attempts because shots made from footage found online have varying numbers of attempts. After gathering data from participants, the proponents will perform feature extraction to gather the points in the shooting motion and soon to be preprocessed. Samples are labeled as free throw shots and three point jumpshots.

C. Feature Extraction

Once the participants are able to perform the shot types asked of them and are captured through video, the proponents will make sure if each video will pass the criteria needed for the study, by making sure if the shooting motion starts with the initial shootingpose and ends with the release of the shot. The frames are extracted from the videos to be used to gather skeletal points. The videos are in 30 fps and when extracted, resulting in the maximum number of frames possible from the video. After the videos are converted into frames, these will be placed into their corresponding participants trial folder. After extracting all the frames, the frames are processed into the program using an OpenPose library. OpenPose outputs the x and y coordinates of every key feature. OpenPose has a problem regarding image noise because OpenPose can detect noises but does not process them as part of the x and y coordinates of the participant as seen in Figure 5. Each frame will be processed and the resulting data would be placed in a csv file. The csv file will contain the following information such as: (1) Indicating the type of shot, (represented as 0 for free throw and 1 for three point shot), (2) If the shot was successful or not (represented as 0 for failed and 1 for successful), (3) trial number and (4) Frames number processed.

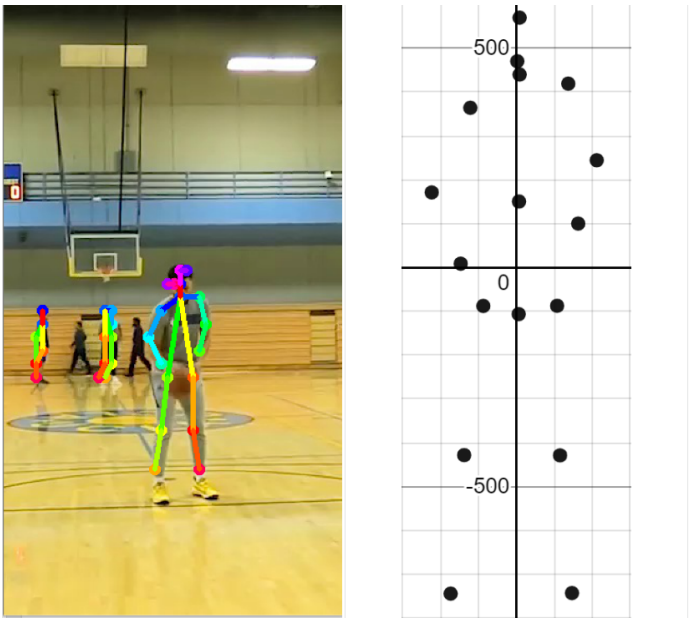


Fig. 5. Noise Detection

D. Preprocessing

1) Data Points: After gathering the skeletal points on every frame from the shooting motion, all frames will undergo affine transformation, which aligns all frames using the x and y coordinates of the skeletal points. The skeletal points will be used to align the frame by splitting the body into two parts, the upper and lower part. The data point of basis for the upper part of the body will be the neck point while the point of basis

for the lower body is the tailbone point. There will be a fixed x and y coordinate value for every frame so that all participant data will be consistently aligned with each other as seen in Figure 6.

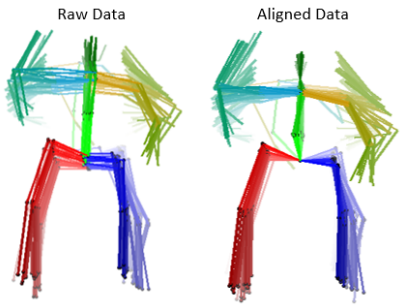
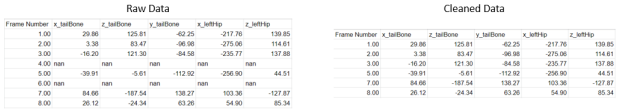


Fig. 6. Affine Transformation

After alignment, the angles and lengths of other significant joints are also computed. From computing lengths and angles, an additional 22 new key features are added.

After performing affine transformation, windowing will be implemented which will allow the data to have a time series representation. The data will be split into consecutive segments (also called windows) to represent the respective time frame of the data, thus have the data presented with a start, middle and end.

The first step to clean the data is to remove irrelevant data from the dataset, such as null values or duplicates, also called NaNs. After removing all NaNs in the dataset, the proponents will check if outliers are present in the dataset. Outliers are needed to be removed for the data to be sure is within the range that it is supposed to be. Windowing is done first before cleaning the dataset so that the time series aspect of the dataset will not be tampered with.



Frame Number	x_tailBone	z_tailBone	y_tailBone	x_leftHip	z_leftHip
1.00	29.86	125.81	-62.25	-217.76	139.85
2.00	3.38	83.47	-96.98	-275.06	114.61
3.00	-16.20	121.30	-84.58	-235.77	137.88
4.00	nan	nan	nan	nan	nan
5.00	-39.91	-5.61	-112.92	-256.90	44.51
6.00	nan	nan	nan	nan	nan
7.00	84.66	-187.54	138.27	103.36	-127.87
8.00	26.12	-24.34	63.26	54.90	85.34

Frame Number	x_tailBone	z_tailBone	y_tailBone	x_leftHip	z_leftHip
1.00	29.86	125.81	-62.25	-217.76	139.85
2.00	3.38	83.47	-96.98	-275.06	114.61
3.00	-16.20	121.30	-84.58	-235.77	137.88
5.00	-39.91	-5.61	-112.92	-256.90	44.51
7.00	84.66	-187.54	138.27	103.36	-127.87
8.00	26.12	-24.34	63.26	54.90	85.34

Fig. 7. Clearning Dataset

Another step to convert the data into time series is to compute for each window's minimum, maximum and average. Computing the three would make the data be more specific in terms of time series. Aside from these, before feeding the dataset to the machine learning models, normalization will be implemented so that the dataset can be read faster by the model and be at the same range.

2) Skeletal Data Image: After cleaning the given set of skeletal points from the gathered video, the sets of segments would be placed together to create one overlayed image. The overlayed image would not be clear because the players actions from the starting frame to the final frame will be compiled together. This will be needed in the study, to be able to represent the changes of coordinates per frame to show the shooting motion of the participant. For example, as seen in Figure 8, it shows a player from start to end motion. The skeletal points with lower opacity represents the start of the shooting motion. Then progressing towards the final frame, in each frame the points and joints change opacity until it reaches to full opacity, representing the change in coordinates per frame, meanwhile the skeletal points with the highest opacity represent the coordinates at the final frame of the shooting motion. The image represents the coordinates of every joint per frame from the shooting motion. The generated image will be used as input for the machine learning models. This will serve as a time series input for the machine learning models because with the overlayed image, it represents the changes in motion at a certain time period. The machine will be able to train and test using this input to check if the player scores or misses the hoop.

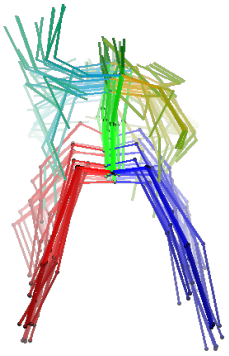


Fig. 8. Overlayed Image of a Person's Shooting Motion

E. Machine Learning

1) Support Vector Machine: Support Vector Machine is a supervised algorithm used for classification and regression. The objective of the support vector machine algorithm is "to find a hyperplane in an N-dimensional space that distinctly classifies the data points", where N is the number of features, which means the dimension of the hyperplane depends on how many features were used. The hyperplanes serve as boundaries in classifying the data. The support vectors in the hyperplane are the data points that are closer to the hyperplane compared to other data points. These support vectors heavily influence the orientation of the hyperplane. The distance of the support vectors determine the model of the support vector machine.

2) Convolutional Neural Networks: Convolutional Neural Network (CNN) is a deep learning algorithm primarily used for image processing. This algorithm is 90% accurate because of its numerous layers: (1) input layer, (2) output layer, (3) hidden layer. Like theother machine learning algorithms, CNN will be used to classify images. The process of the training using CNN is forward propagation and backward propagation. It repeats this step until it is trained to detect the correct form of the player. CNN will be the algorithm to be used for deep learning of the generated image.

After preprocessing the data, machine learning would be implemented to analyze the different shooting forms if the shot is made or not using both the preprocessed time series data and the generated image. The collected data would be split using two different methods such as user fold validation and cross fold validation. User fold validation would isolate one participant's trial, which would be the test, from the rest of the dataset while cross fold validation would split train to 70% and test to 30%. The collected data would be separated from professionals and non-professionals, creating a separate model for each group, in order to have a more accurate analysis since professionals have different skill sets compared to non-professionals. Models are also separated for each shot type since each shot type has different difficulty, which are: Free Throw Shot and Three Point Shot. The trained dataset would be divided into two, supervised learning and unsupervised learning. The test data will be given to the machine to predict if the player scored the shot or missed the shot. After training the trained dataset, each machine learning model used in predicting the shot would be evaluated through the following performance metrics: accuracy, precision, recall, f1 score, and confusion matrix.

IV. Results

Due to the lack of skeletal data gathered from the professional players performing free throw shots, the data gathered is significantly unbalanced having 16 successful shots while only having 2 failed shots. The researchers tried inputting the data through the SVM and CNN models, where the results yielded unbalanced results, where the player that had 2 failed shots are being classified by the model as a failed shot even if it is a successful shot since there were no trained successful shots of the player.

A. Support Vector Machine - Cross Fold Validation

Accuracy = 80.9524%	**Precision**	**Recall**	**F1 Score**
0.0 (shot not made)	0.82	0.82	0.82
1.0 (shot is made)	0.80	0.80	0.80

TABLE III
Results of SVM for Non-Professional Free Throw Shot Model

	0.0	**1.0**
0.0	9	2
1.0	2	8

TABLE IV
Confusion Matrix for Non-Professional Free Shot Model

1) Non-Professional Free Throw Shot: As seen in Table III, the SVM model for the non professionals' free throw dataset has an accuracy of around 81% under supervised learning. This indicates that with the model using 70% of the non-professionals free throw dataset and testing 30% of the said dataset, it was able to correctly classify more than 80% of the testing set. With a cross validated total of 11 samples for the 0.0 label and 10 samples for the 1.0 label, the model is tested with two labels indicating if a shot is made or not, 0.0 for the former and 1.0 for the latter. This can be seen in Table IV, where 0.0 is the positive value and 1.0 is the negative value, the model has a True Positive value of 9, False Positive value of 2, False Negative Value of 2 and True Negative value of 8. The precision score for the 0.0 label has yielded a 0.82 score which indicates that among all classifications that are predicted as missed shots, around 82% of the 0.0 classifications are correctly classified. This shows that the model is precise in classifying missed free throws because it pretty much exceeds more than 0.5 of the total classified as missed free throws. The precision score for the 1.0 label has yielded a 0.80 score, 0.02 less than the precision score of the 0.0 label. This is still a precise classification result because it has also correctly classified 80% of all 1.0 classifications has only 0.02 difference than the precision score of the 0.0 label. Similarly, the recall scores for both the 0.0 and 1.0 labels are the same with their precision scores respectively. This indicates that among all classifications that are actually missed shots, around 82% of the 0.0 label is correctly classified. Also with the 1.0 label, its recall score indicates that among all classifications that are actually made shots, 80% of the 1.0 label is correctly classified. This shows that the model has high recall scores for both labels. With both labels having the same precision and recall scores, the F1 score for both labels are the same, where the 0.0 label has 0.82 and the 1.0 label has 0.80 respectively.

Accuracy = 81.8182%	**Precision**	**Recall**	**F1 Score**
0.0 (shot not made)	0.80	0.80	0.80
1.0 (shot is made)	0.83	0.83	0.83

TABLE V
Results of SVM for Non-Professional Three Point Shot Model

	0.0	**1.0**
0.0	4	1
1.0	1	5

TABLE VI
Confusion Matrix for Non-Professional Three Point Shot Model

2) Non-Professional Three Point Shot: The results for the SVM model for the non-professional three point shot dataset, which can be seen in Table V, shows that similar to the non professional free throw shot model, it has around 82% accuracy under supervised learning. This model also has a cross validation dataset split of 70% training and 30% testing. As seen in Table VI, where 0.0 is the positive value and 1.0 is the negative value, the model has True Positive value of 4, False Positive value of 1, False Negative value of 1, and True Negative value of 5. This model can be seen as an accurate model because it correctly classified more than 80% of the testing set. This model is quite similar to the non professional free throw shot model because both labels' precision and recall scores are the same. The 0.0 label has a precision and recall score of 0.80, which indicates that the classifications of 0.0

labels are 80% correct when among classifications predicted as missed shots and among classifications that are actually missed shots. The 1.0 label has a precision and recall score of 0.83, which states that 83% of 1.0 classifications are correctly classified among classifications predicted as made shots and actual made shots. The F1 score for the 0.0 label and 1.0 label are 0.80 and 0.83 respectively.

Accuracy = 77.7778%	**Precision**	**Recall**	**F1 Score**
0.0 (shot not made)	0.90	0.75	0.82
1.0 (shot is made)	0.62	0.83	0.71

TABLE VII
RESULTS OF SVM FOR PROFESSIONAL THREE POINT SHOT MODEL

	0.0	**1.0**
0.0	9	3
1.0	1	5

TABLE VIII
CONFUSION MATRIX FOR PROFESSIONAL THREE POINT SHOT MODEL

3) Professional Three Point Shot: For the Support Vector Machine model for professional three point shot, as seen in Table VII, it has garnered an accuracy of around 78%, around 4% less compared to the non professional three point shot model. This indicates that this model has more incorrect classifications compared to the other SVM models. As seen in Table VIII, the True Positive value is at 9, the False Positive value at 3, the False Negative value at 1 and the True Negative value at 5. It is evident when observing the precision score for the 1.0 labels and the recall score of the 0.0 labels, which are 0.62 and 0.75 respectively. This means that only 62% of 1.0 labels are correctly classified among all the made shots classifications, while only 75% of 0.0 labels are correctly classified among classifications that are actually made shots. These scores are lower compared to the other models which have scores around 0.80. These are likely due to the model being slightly overfitted because when samples are fed into the model, there are more samples in the 0.0 label than the 1.0 label in the training set and the testing set, with the latter having samples of 12 and 6 respectively. These are also due to datasets of online participants where some labels can have more trials compared to the other due to these participants' footage being limited with the internet's resources. However, due to these, the precision score of the 0.0 label is high at around 0.90 and the recall score of the 1.0 label is at 0.83, which is also a good recall score. The F1 scores for the 0.0 label and 1.0 label are at 0.82 and at 0.71, still far above 0.5. Despite having some lower precision and recall scores compared to the previous models, this model still has yielded good accuracy and F1 scores where are still far above 0.5.

B. Support Vector Machine - User Fold Validation

User ID	Accuracy	Precision	Recall	F1 Score	Kernel
0	60.00%	0.60	0.62	0.58	Sigmoid
1	30.00%	0.15	0.50	0.23	Poly
2	30.00%	0.30	0.29	0.29	Sigmoid
3	60.00%	0.62	0.62	0.60	Sigmoid
4	80.00%	0.85	0.80	0.79	Sigmoid
5	60.00%	0.58	0.58	0.58	Linear
6	90.00%	0.87	0.92	0.89	Sigmoid
7	40.00%	0.37	0.37	0.37	Sigmoid

TABLE IX
RESULTS OF SVM FOR THE USER FOLD VALIDATION NON-PROFESSIONAL FREE THROW SHOT DATASET

User ID		**0.0**	**1.0**
0	**0.0**	2	1
	1.0	3	4
1	**0.0**	3	0
	1.0	7	0
2	**0.0**	2	4
	1.0	3	1
3	**0.0**	3	1
	1.0	3	3
4	**0.0**	3	2
	1.0	0	5
5	**0.0**	4	2
	1.0	2	2
6	**0.0**	3	0
	1.0	1	6
7	**0.0**	1	3
	1.0	3	3

TABLE X
CONFUSION MATRIX OF SVM FOR THE USER FOLD VALIDATION NON-PROFESSIONAL FREE THROW SHOT DATASET

1) Non-Professional Free Throw Shot: For the user fold validated non-professional free throw shot SVM model, as seen in Table IX, three kernels used for testing are Poly, Linear and Sigmoid. For the Poly kernel result, which is participant 1, it can be seen that using this kernel resulted to low accuracy, recall and precision, where as seen in its confusion matrix at Table X, the model only had correct classifications on missed shots (0.0) and had all classifications for made shots (1.0) incorrect as seen in its False Negative and True Negative values at 7 and 0. For the Linear kernel result, it can be seen that it yielded a 60% accuracy. Most of the other tests used Sigmoid as a kernel, which showed both better and worse results. The accuracy for these results range from 30 to 90%, with participant 2 and 7 having both lowest accuracy at 30% and 40%. Both have large proportions of wrong classifications on both missed and made shots as seen in the False Positive and False Negative values on both confusion matrices at Table X. This can be related to why the combined precision, recall, and f1 scores of the two labels for both tests are low, ranging from 0.15 to 0.50. This is due to the model not recognizing these shots features having similar data compared to the data included in the training dataset, even with using the best possible kernel for testing. However, despite having two results having bad performances, most tests under Sigmoid have above 50% accuracy, ranging from 60% to the best 90%.

Tests having 60% accuracy, which are participant 0 and 3, can be observed having precision, recall and f1 scores near 0.60. This can be seen in their respective confusion matrices that both have higher False Negative values than and only having False Positive values at 1. This is due to the model having somewhat similar features with others on the training set while other features in the training set the model is not familiar with. It can be observed that participant 6 has a high 90% accuracy, as seen in its False Negative value at 1, out of all classifications of made shots, only 1 of 7 classifications is incorrect. Participant 4 is also similar, having an 80% accuracy. This test has a 0 False Negative value, while having a 2 False Positive value. Both tests have high accuracy, precision, recall and f1 scores because the model recognized the features of this shooting form mostly making shots.

User ID	Accuracy	Precision	Recall	F1 Score	Kernel
0	60.00%	0.60	0.62	0.58	Sigmoid
1	60.00%	0.60	0.62	0.58	Sigmoid
2	30.00%	0.30	0.29	0.29	Sigmoid
3	60.00%	0.62	0.62	0.60	Sigmoid
4	80.00%	0.85	0.80	0.79	Sigmoid
5	60.00%	0.58	0.58	0.58	Sigmoid
6	90.00%	0.87	0.92	0.89	Sigmoid
7	40.00%	0.37	0.37	0.37	Sigmoid

TABLE XI
RESULTS OF SVM USING SIGMOID KERNEL FOR THE USER FOLD VALIDATION NON-PROFESSIONAL FREE THROW SHOT DATASET

User ID		**0.0**	**1.0**
0	**0.0**	2	1
	1.0	3	4
1	**0.0**	2	1
	1.0	3	4
2	**0.0**	2	4
	1.0	3	1
3	**0.0**	3	1
	1.0	3	3
4	**0.0**	3	2
	1.0	0	5
5	**0.0**	4	2
	1.0	2	2
6	**0.0**	3	0
	1.0	1	6
7	**0.0**	1	3
	1.0	3	3

TABLE XII
CONFUSION MATRIX OF SVM USING SIGMOID KERNEL FOR THE USER FOLD VALIDATION NON-PROFESSIONAL FREE THROW SHOT DATASET

To have a more consistent observation and consensus of the Non Professional Free Throw shot model, the Sigmoid kernel is used for each testing sample to see if the metrics change under different kernels. For participant 1, initially using the Poly kernel, the metrics scores saw an increase in accuracy, precision, recall and F1 score, with now having the exact same metrics scores with participant 0. It is evident in the confusion matrix, wherein the True Positive and True Negative values increased significantly from initially both having 0. This indicates that these two samples have similar features when using the same kernel, not seen in the previous observation. For participant 5, there are no changes on the results for this sample when it was changed from Linear kernel to Sigmoid kernel. This further shows that the Non-Professional Free Throw model has higher results.

User ID	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Kernel**
0	30.00%	0.61	0.56	0.29	Linear
1	60.00%	0.33	0.43	0.38	Linear
2	50.00%	0.28	0.42	0.33	Linear
3	44.44%	0.42	0.42	0.42	Linear
4	40.00%	0.38	0.38	0.38	Sigmoid
5	50.00%	0.64	0.69	0.49	Sigmoid

TABLE XIII
RESULTS OF SVM FOR THE USER FOLD VALIDATION NON-PROFESSIONAL THREE POINT SHOT DATASET

User ID		**0.0**	**1.0**
0	**0.0**	2	0
	1.0	7	1
1	**0.0**	6	1
	1.0	3	0
2	**0.0**	5	1
	1.0	4	0
3	**0.0**	3	3
	1.0	2	1
4	**0.0**	3	3
	1.0	3	1
5	**0.0**	2	0
	1.0	5	3

TABLE XIV
CONFUSION MATRIX OF SVM FOR THE USER FOLD VALIDATION NON-PROFESSIONAL THREE POINT SHOT DATASET

2) Non-Professional Three Point Shot: The results of the user fold validated non professional three point shots, seen in Table XIII, shows that four out of six tests used a Linear kernel, while the other two are using Sigmoid kernel. This model, with both the Linear and Sigmoid kernel, yielded low results for accuracy, with the highest being 60% and the lowest at 30%. With only one test having an accuracy higher than 0.5 indicates that the model cannot classify and distinguish the data properly. It can be seen that the precision, recall and f1 scores are also not high, indicating that various proportions on the combined value of true positives and true negatives are not high, making the classifications of each test not consistent. This is due to the model not interpreting a pattern among the features given in the training set to be able to consistently classify the shot. The three points shot is also somewhat dependent on the flick of the wrist, not covering the entire movement of the hand. The distance towards the ring should be considered since three point shots have a lower probability of being made and are unpredictable compared to the free throw.

User ID	Accuracy	Precision	Recall	F1 Score	Kernel
0	30.00%	0.61	0.56	0.29	Sigmoid
1	60.00%	0.33	0.43	0.38	Sigmoid
2	50.00%	0.28	0.42	0.33	Sigmoid
3	44.44%	0.42	0.42	0.42	Sigmoid
4	40.00%	0.38	0.38	0.38	Sigmoid
5	50.00%	0.64	0.69	0.49	Sigmoid

TABLE XV
RESULTS OF SVM USING SIGMOID KERNEL FOR THE USER FOLD VALIDATION NON-PROFESSIONAL THREE POINT SHOT DATASET

User ID		0.0	1.0
0	**0.0**	2	0
	1.0	7	1
1	**0.0**	6	1
	1.0	3	0
2	**0.0**	5	1
	1.0	4	0
3	**0.0**	3	3
	1.0	2	1
4	**0.0**	3	3
	1.0	3	1
5	**0.0**	2	0
	1.0	5	3

TABLE XVI
CONFUSION MATRIX OF SVM USING SIGMOID KERNEL FOR THE USER FOLD VALIDATION NON-PROFESSIONAL THREE POINT SHOT DATASET

To have a more consistent observation and consensus of the Non Professional Three Point shot model, all samples used the Sigmoid kernel. As seen from Participant 0 to 3, all the metrics results for these samples did not change, having the same results as the previous observation, where Linear is used as the kernel. This shows that the consensus for the model is the same as the previous one because no results have changed.

User ID	Accuracy	Precision	Recall	F1 Score	Kernel
0	50.00%	0.50	0.25	0.33	Sigmoid
1	57.14%	0.62	0.75	0.53	Linear
2	33.33%	0.35	0.33	0.33	Sigmoid
3	66.66%	0.43	0.38	0.40	Sigmoid
4	70.00%	0.67	0.69	0.67	Sigmoid
5	55.55%	0.53	0.54	0.50	rbf
6	40.00%	0.67	0.57	0.38	Sigmoid
7	80.00%	0.50	0.40	0.44	Sigmoid
8	62.50%	0.63	0.62	0.62	Sigmoid
9	50.00%	0.47	0.47	0.47	Sigmoid
10	60.00%	0.50	0.30	0.37	Sigmoid

TABLE XVII
RESULTS OF SVM FOR THE USER FOLD VALIDATION PROFESSIONAL THREE POINT SHOT DATASET

User ID		0.0	1.0
0	**0.0**	0	0
	1.0	2	2
1	**0.0**	1	0
	1.0	3	3
2	**0.0**	1	2
	1.0	4	2
3	**0.0**	0	1
	1.0	2	6
4	**0.0**	5	2
	1.0	1	2
5	**0.0**	1	1
	1.0	3	4
6	**0.0**	3	0
	1.0	6	1
7	**0.0**	0	0
	1.0	1	4
8	**0.0**	2	2
	1.0	1	3
9	**0.0**	1	2
	1.0	2	3
10	**0.0**	0	0
	1.0	4	6

TABLE XVIII
CONFUSION MATRIX OF SVM FOR THE USER FOLD VALIDATION PROFESSIONAL THREE POINT SHOT DATASET

3) Professional Three Point Shot: As seen in the Table XVII, the results of the user fold validated professional three point shots primarily have Sigmoid as the kernel. Sigmoid comes from a neural network thus being more powerful than linear or rbf. The dataset for this category might be all over the place thus leading it to need a more powerful kernel to train the dataset. It can be observed in the results that the accuracy for each test varies, ranging from 33% to 70%. The result of participant 7, which is at a high 80%, will not be discussed due to not having samples for missed shots and it is evident with the performance metrics having low scores due to the unbalanced samples. For the tests having a below 0.5 accuracy, it can be observed that the values of False Positives and False Negatives are high as seen with their confusion matrices at Table XVIII. This causes the results' F1 score and accuracy to be low, showing that the model does not recognize the forms' features to determine if the shot is made or not. Looking into the results that have accuracies ranging from 50% to 70%, these results have precision, recall and F1 scores ranging from 0.47 to 0.69. This is determined by the size of the test samples and how much the model classified correctly. Some have lower scores due to having a sizable proportion of wrong classifications while others have better scores because of its proportion on the higher edge. It could be noticed that participant 3's accuracy result is high albeit it's low precision, recall and F1 score. This is because as seen in its confusion matrix, the number of testing samples for missed (0.0) shots is at 1, compared to the made shots (1.0) at 8 samples, also placed at the False Positive value thus lowering the metrics' values. Lastly, the best result, which participant 4's accuracy result at 70%, only has small values at False Positive and False Negative. This result did not have a higher accuracy due to the testng sample of made shots (1.0) having a small testing sample. Since in comparison to the non professional

model, the best result from this model (which is at 70%) is higher compared to the previous one is, despite also putting distance as a factor, because professional players have more consistent forms and shot results when shooting, indicating that the level of play of a player can be noticed by the model's classifications.

User ID	Accuracy	Precision	Recall	F1 Score	Kernel
0	50.00%	0.50	0.25	0.33	Sigmoid
1	42.85%	0.38	0.25	0.30	Sigmoid
2	33.33%	0.35	0.33	0.33	Sigmoid
3	66.66%	0.43	0.38	0.40	Sigmoid
4	70.00%	0.67	0.69	0.67	Sigmoid
5	55.55%	0.53	0.54	0.50	Sigmoid
6	40.00%	0.67	0.57	0.38	Sigmoid
7	80.00%	0.50	0.40	0.44	Sigmoid
8	62.50%	0.63	0.62	0.62	Sigmoid
9	50.00%	0.47	0.47	0.47	Sigmoid
10	60.00%	0.50	0.30	0.37	Sigmoid

TABLE XIX
RESULTS OF SVM USING SIGMOID KERNEL FOR THE USER FOLD VALIDATION PROFESSIONAL THREE POINT SHOT DATASET

User ID		0.0	1.0
0	**0.0**	0	0
	1.0	2	2
1	**0.0**	0	1
	1.0	3	3
2	**0.0**	1	2
	1.0	4	2
3	**0.0**	0	1
	1.0	2	6
4	**0.0**	5	2
	1.0	1	2
5	**0.0**	1	1
	1.0	3	4
6	**0.0**	3	0
	1.0	6	1
7	**0.0**	0	0
	1.0	1	4
8	**0.0**	2	2
	1.0	1	3
9	**0.0**	1	2
	1.0	2	3
10	**0.0**	0	0
	1.0	4	6

TABLE XX
CONFUSION MATRIX OF SVM USING SIGMOID KERNEL FOR THE USER FOLD VALIDATION PROFESSIONAL THREE POINT SHOT DATASET

To have a more consistent observation and consensus of the Professional Three Point shot model, the Sigmoid kernel was used for all samples. As seen in the results of Participant 1, which previously used the Linear kernel, all the metrics results decreased. From having an 57.14% accuracy, 0.62 precision, 0.75 recall and 0.53 F1 score, it decreased to 42.85% accuracy, 0.38 precision, 0.25 recall and 0.30 F1 score. The Sigmoid kernel was not able to correctly classify the 1 0.0 classification, as seen in the confusion matrix of Participant 1 having a False Positive value of 1. This resulted in lower results for Participant 1. Participant 5, using the rbf kernel in the previous observation, still has the same results when converted to the Sigmoid kernel. Albeit having one testing with a lower result compared to the previous observation due to using Sigmoid as the kernel for all testing samples, the general consensus of the model is still the same due to the decreased result still being within range of the highest and lowest accuracies.

C. Convolutional Neural Network - Cross Fold Validation

Epoch	Loss	Accuracy	Precision	Recall	F1 Score
1	3.3956	0.6000	0.6364	0.6364	0.6364
2	0.6217	0.7143	0.7857	0.7857	0.7857
3	0.5143	0.7500	0.6923	0.9000	0.7826
4	0.6429	0.8000	0.8750	0.7000	0.7778

TABLE XXI
RESULTS OF CNN FOR THE NON-PROFESSIONAL FREE THROW DATASET WITH EPOCH

1) Non-Professional Free Throw Shot: As seen in Table XXI, the models for the CNN were trained and tested in four epochs. For the CNN model for non professional free throws, it can be observed that at the first epoch, the accuracy is at 0.6000, which is very low, just barely above 0.5. Also during the first epoch, both the precision, recall and F1 score of the model are also low, all three scoring only at 0.6364, which are also barely above 0.5. This indicates that only 63.64% of classifications are correct among the classes' actual and predicted classifications, showing that the model is not precise. For the second epoch, the accuracy of the model is at 0.7143, or 71.43%, higher than the previous epoch by 0.1143. This epoch also showed an increase in precision, recall and F1 score, all three at 0.7857. For the third epoch, the results yielded by the model have again improved compared to the previous epoch. The model's accuracy increased to 0.7500. Despite having a higher recall score of 0.9000, the precision score is lower compared to the previous epoch at 0.6923, resulting in the model having an F1 score of 0.7826. With the final epoch of this model, it is similar with the previous epoch, where the model became most accurate having an accuracy of 0.8000. It also had the highest precision score at 0.8750. However, the model's loss increased at 0.6429 and the model's recall score decreased mildly at 0.7000, resulting in an F1 score of 0.7778. With this model, it shows that with each increasing epoch, the more accurate and precise the model became, albeit some metric scores decreasing. It can be observed from the precision, recall and F1 score from each epoch ranges from 0.6000 to 0.9000, showing that features from the training dataset are recognized by the model with some shooting forms but due to varying data points across different shooting forms, the model can hardly classify the shot made. With each epoch, the model learns from the different data points and angles, increasing its accuracy.

Epoch	Loss	Accuracy	Precision	Recall	F1 Score
1	8.2118	0.6000	0.7500	0.7500	0.7500
2	0.6503	0.6667	0.5556	0.6250	0.5883
3	0.6255	0.7500	0.8750	0.6364	0.7369
4	0.3643	0.8500	0.8182	0.9000	0.8572

TABLE XXII
RESULTS OF CNN FOR THE NON-PROFESSIONAL THREE POINT SHOT DATASET WITH EPOCH

2) Non-Professional Three Point Shot: For the CNN model for non professional three point shots, as seen in Table XXII. It can be observed that the trajectory of the results is similar with the previous model. At the first epoch, the accuracy of the model is only at 0.6000, only 60% accurate, only slightly above 0.5. The model's precision, recall and F1 score are at 0.75, which are both good scores but the accuracy is not high enough. The loss for this model is also high, similar to the previous epoch, at 8.2118, indicating a high prediction error. Improvement can be observed with the following epochs. By the second epoch, accuracy increased slightly at 0.6667, however the precision, recall and F1 score of the model decreased into 0.5556, 0.6250 and 0.5883 respectively. The loss of the model decreased at 0.6503. For the third epoch, the model's results can be seen improving compared to the second epoch, the accuracy is at 0.7500, the precision is at 0.8750, the recall is at 0.6364, and the F1 score is at 0.7369. In the last epoch, the model is able to yield its highest accuracy at 0.8500 and its highest recall score at 0.9000, and its highest F1 score at 0.8572. Despite the precision getting lower by a few at 0.8182, the model became more precise and accurate for each epoch. Similar to the non professional free throw dataset, the pattern in which the accuracy's improvement with each epoch, the model learns from the various features included with the dataset. However, it could be observed that this model has lower results compared to the non professional free throw data set because the distance of the participant to the ring is farther when shooting free throw shots, having a smaller probability in making shots, making this shot type more unpredictable.

Epoch	Loss	Accuracy	Precision	Recall	F1 Score
1	5.5338	0.6296	0.7143	0.7895	0.7500
2	0.5902	0.6667	0.7391	0.8500	0.7907
3	0.3070	0.8148	0.8400	0.9545	0.8936
4	0.4698	0.8148	0.8500	0.8947	0.8718

TABLE XXIII
RESULTS OF CNN FOR THE PROFESSIONAL THREE POINT SHOT DATASET WITH EPOCH

3) Professional Three Point Shot: The CNN model for professional three point shots, as seen in Table XXIII. It can be seen that the results are similar with the previous models above. The first epoch yielded an accuracy of 0.6296, roughly 63%. The model got a precision of 0.7143, a recall of 0.7895, and an F1 score of 0.8125, which is a good score. However, the loss of the model is high at 5.5338 which indicates a high prediction error. The following epoch shows slight improvement with the accuracy going up to 0.6667, precision went up to 0.7391, recall to 0.8500, and F1 score at 0.7907, while the loss for the model went down to 0.5902. For the third epoch we can still observe the slight improvement from the previous epoch with the accuracy going up to 0.8148, precision to 0.8400, recall to 0.9545, F1 score to 0.8936. The loss also went down to 0.3070. However for the fourth epoch, the accuracy stays at 0.8148, the precision went up to 0.8500, recall went down to 0.8947, the F1 score also went down at 0.8718 and the loss of the model went up to 0.4698.

D. Convolutional Neural Network - User Fold Validation

User ID	Loss	Accuracy	Precision	Recall	F1 Score
0	0.6585	0.6316	0.7143	0.5000	0.5882
1	0.5974	0.7143	0.6364	0.7778	0.7000
2	0.6481	0.7619	0.6875	1.0000	0.8148
3	0.5351	0.7619	0.7619	1.0000	0.8649
4	0.6398	0.6667	0.6875	0.8462	0.7586
5	0.5854	0.7619	0.7059	1.0000	0.8276
6	0.6352	0.7143	0.6000	0.7500	0.6667
7	0.5440	0.8824	0.9091	0.9091	0.9091

TABLE XXIV
RESULTS OF CNN FOR THE USER FOLD VALIDATION NON-PROFESSIONAL FREE THROW SHOT DATASET

1) Non-Professional Free Throw Shot: For the user fold validated CNN models of non professional free throw shots shown at Table XXIV, the accuracy for each participant ranges from 0.6316 to 0.8824. Some of the highest results range around 0.7143 to 0.8824. Among all results, it can be seen that by the fourth epoch, each result does not have a very bad loss value, around 53-66%. It can be observed that participant 7 has the highest accuracy, precision, recall, and F1 score, showing that the model recognized most of this participant's features, with only an estimate of 10% of false positive and false negative values. The results that have an accuracy of 0.7619 can be observed with having recall scores of 1.000. This indicates that these results produced no false negatives, meaning that it did not have wrong classifications on made shots (1.0). This shows that the model having trained under these results features similar to their respective shooting forms, recognizes them when making shots, not missing. This could be improved by adding more random shot data. The results having lower accuracy, which are around 0.6316 and 0.6667, can be noticed for having some precision and recall scores ranging from 0.5000 to 0.7000. This indicates that these results have a larger number of false positives or false negatives compared to the results. This is due to their features not recognized by the model due to these features are not that familiar among the training set. Distance can be a factor because free throw shots are made more frequently due to having a closer distance to the ring, having a higher probability in making.

User ID	Loss	Accuracy	Precision	Recall	F1 Score
0	0.4579	0.8095	0.7500	0.7500	0.7500
1	0.5231	0.8095	0.7500	0.9000	0.8182
2	0.4356	0.8095	0.8333	0.8333	0.8333
3	0.7701	0.6500	0.6667	0.6000	0.6316
4	0.7585	0.6000	0.5000	0.6250	0.5556
5	0.5260	0.8095	0.7500	0.7500	0.7500

TABLE XXV
RESULTS OF CNN FOR THE USER FOLD VALIDATION NON-PROFESSIONAL THREE POINT SHOT DATASET

2) Non-Professional Three Point Shot: For the user fold validated CNN models of non professional three point shots, the model ran for four epochs and the results presented are of the fourth and last epoch. The six participants used for testing

showed both similar and different results from different metrics. It can be observed that some participants have a somewhat low accuracy and some have good accuracy rates, ranging from 0.6000 to 0.8095. When observing each participant's accuracy, four of the six total accuracy results are at 0.8095 while two are at 0.6500 and 0.6000 respectively. The precision, recall and F1 scores for the results with good accuracy are ranging from 0.7500 to 0.9000 while the results with lower accuracy are ranging from 0.5000 to 0.6667. The results of participants with good accuracy have these scores of precision, recall and F1 score is because when each participant is used for testing with the rest of the participants used for training, this indicates that around 75-90% of the classifications of these participants have mostly similar features that of the training set. This is likely that these participants' shooting forms have a similar set of skeletal data points and angles on important joints when making or missing shots that made the model classify most of it correctly, only having around 10-25% of false positive and false negative values on the testing set. The results with lower accuracy, precision and recall have these results because the features for these are not that similar with the others. It can be seen from the false positive and false negative values of these results are at 37.50% to 50%, indicating that the skeletal data points and angles of these participants are slightly different from the other participants, making the model having wrong classifications for these. Distance is a factor on the classifications due to the farther the distance, the shot being more unpredictable if it is made or not.

User ID	Loss	Accuracy	Precision	Recall	F1 Score
0	0.6292	0.7619	0.7619	1.0000	0.8649
1	0.6783	0.4762	0.5556	0.7692	0.6452
2	0.6271	0.7143	0.7000	1.0000	0.8235
3	0.6235	0.8571	0.8571	1.0000	0.9231
4	0.6390	0.6667	0.6667	1.0000	0.8000
5	0.5325	0.8095	0.8000	1.0000	0.8889
6	0.5617	0.8095	0.8000	1.0000	0.8889
7	0.2931	0.9524	0.9524	1.0000	0.9756
8	0.6716	0.6667	0.6667	1.0000	0.8000
9	0.7129	0.5789	0.6111	0.9167	0.7333
10	0.7022	0.6667	0.7647	0.8125	0.7879

TABLE XXVI
RESULTS OF CNN FOR THE USER FOLD VALIDATION PROFESSIONAL THREE POINT SHOT DATASET

3) Professional Three Point Shot: For the user fold validated CNN models of professional three point shots, the model ran for four epochs and the results presented are of the last epoch. It can be seen that the accuracy of the participants ranges from the lowest at 0.4762 to the highest at 0.9524. Meanwhile the recall and F1 score of the lowest accuracy also has the lowest recall and F1 score compared to all of the participants. While those with good accuracy have a recall of 0.7692 to 1.000. Eight of the participants' recall has a value of 1.000 which means that these participants did not have false negative values. Moreover, the recall of the participants is high ranging from 0.5556 to 0.9524. The precision scores of the participants vary, ranging from 0.5556 to 0.8571, all above 0.5. With good recall and precision, most of the tests have a good proportion of True Positive and True Negatives, having more correct classifications. It can be observed here that 5 of the 11 participants have the same accuracy and precision.The lowest loss value comes from the participant that has the highest accuracy which is at 0.2931 while the other participants have a loss value ranging from 0.5325 to 0.7129, indicating that the model does not have a very bad performance in classification. This is also evident in the F1 scores generated by the model, with only the lowest at 0.6452 and the highest F1 score at 0.9756. This model has a higher accuracy than the previous model (non profession three point shot), which shows that model can recognize the professional shot type better in classification.

V. SUPPORT VECTOR MACHINE VS CONVOLUTIONAL NEURAL NETWORK

Dataset	Accuracy	Precision	Recall	F1 Score
Non-Professional Free Throw	60.00%	0.59	0.60	0.54
Non-Professional Three Point Shot	45.74%	0.44	0.48	0.38
Professional Three Point Shot	55.53%	0.51	0.43	0.45

TABLE XXVII
SUMMARY OF RESULTS OF SVM FOR THE USER FOLD VALIDATION

Dataset	Accuracy	Precision	Recall	F1 Score
Non-Professional Free Throw	73.37%	0.70	0.84	0.77
Non-Professional Three Point Shot	74.16%	0.70	0.74	0.72
Professional Three Point Shot	71.71%	0.73	0.95	0.38

TABLE XXVIII
SUMMARY OF RESULTS OF CNN FOR THE USER FOLD VALIDATION

To summarize all the results gathered from all the models, the proponents gathered all the results from each metric and computed the average to give a consensus. The user fold validation models are used for the final observation because user fold validation is able to observe common features that factors a shot result in a given set of data with each other compared to randomly selected data in feature extraction, showing the results for user fold validation is more consistent. For the SVM user fold validated Non-Professional Free Throw models, the average accuracy is at 60%, precision at 0.59, and recall at 0.60. For the SVM user fold validated Non-Professional Three Point Shot models, the average accuracy is at 45.75%, precision at 0.44, and recall at 0.48. For the SVM user fold validated Professional Three Point Shot models, the

average accuracy is at 55.53%, precision at 0.51, and recall at 0.43. For the CNN user fold validated Non-Professional Free Throw models, the average accuracy is at 73.37%, precision at 0.70, and recall at 0.84. For the CNN user fold validated Non-Professional Three Point Shot models, the average accuracy is at 74.16%, precision at 0.70, and recall at 0.74. For the CNN user fold validated Professional Three Point Shot models, the average accuracy is at 71.71%, precision at 0.73, and recall at 0.95. When comparing the results of the Support Vector Machine Models and the Convolutional Neural Network models, it can be observed that the CNN models results are higher compared to the SVM models as seen in their respective Accuracy, Precision, Recall and F1 score results. As explained above, familiarity of the features, trajectory of the shot towards the ring, camera angles, level of play and distance are factors on misclassifications. With each average accuracy of the CNN models estimated between 70-75% compared to the SVM models' average accuracy ranging at 45-60%, this indicates that the CNN models are more accurate compared to the SVM models. Another metric to be observed is the F1 scores of both types due to the F1 score being the representational value that balances the variables observed in both precision and recall. The average F1 scores of all SVM models are at 0.54, 0.38, and 0.45 while the average F1 scores of all CNN models are at 0.77, 0.72 and 0.83, indicating that the CNN models are the better models. However in this study, the SVM models, although having lower results, can be seen as more realistic compared to the CNN results because the study focused more on the application and implementation of SVM compared to CNN, indicating that the results of the SVM models are more sensible and reasonable.

VI. Conclusion

With the proponents' observations in the study, it can be observed in the results of the study that skeletal data applied in a time series preprocessing method to be trained and tested to classify various shot types is feasible. However, it is not consistent and accurate enough through different test results, so it can be concluded that different features, different time series methods and different angles and others could be tried and tested to improve the different machine learning models used in the study. Presenting the data in time series is shown to be effective, wherein it informs the model of the data points movement in each corresponding frame, showing the relation of time in studying a player's shot form and release. The inputs for the different features to be observed in the study, both the preprocessed skeletal data and the generated image, is shown that it can be used for observation in machine learning.

To summarize the results of the user fold validated SVM models, the accuracy for varies. For the user fold validated SVM, it can be seen that the free throw yielded higher accuracy results compared to three point shots. The best hyperplane achieved in testing the free throw set is at an 80-90% accuracy, with f1 scores at 0.79 and 0.89. The three point shot models have lower accuracy but it is evident that the professional three point shot model yielded better results than the non-professional. The level of play among players could be a factor since professionals shoot the ball more consistently. It can be observed from the user fold validated models, that the False Positive and False Negative values of each model that some classifications fail may be due to the similarity of the various shooting forms and only the forms' features of the participants are being observed, not the direction of the participant's trajectory in relation to the ring. The model may interpret the shooting forms as perfect but due to the trajectory or direction the participant is facing, it causes some of the classifications to fail. For three point shots, the distance should be considered because farther distances lessen the probability of the shot. Also to be considered are the other skeletal points in the hand beyond the wrist because three point shots can be dependent on the flick of the wrist.

To summarize the results of the user fold validated CNN models, most of the scores yielded by the models can be observed that the precision increases drastically from the first epoch run up to the last epoch run. The professional three point model did not increase in the last epoch, but remained the same with the third epoch. However, the precision and recall values vary, with some having low scores while others have good scores. This is likely due to the testing sample size and the models familiarity with the labels, having some classify correctly while some do not. The proponents also concluded that due to the differences of the angles from the Professional models, some of the data varied due to the placement of the camera when the shot was taken, resulting in insignificant trends in the accuracy and precision scores. Results in the models gave the proponents an idea that distance could be a factor that some classifications are making the model unpredictable. This can be observed in the Non Professional Three Point shots model. Some resulted with a lower accuracy after the last epoch run compared to the others. This indicated that some skeletal data points and angles of each participant are slightly different from other participants.

VII. Recommendation

After the conclusion of the study, the proponents recommend to try other classification machine learning algorithms such as Decision Trees, Naive Bayes Classifier, Logistic Regression and K-Nearest Neighbors where it is possible for these other classification algorithms to be more accurate. It is also recommended to try other models of CNN such as Inception and ResNet. Jump shots for each shot type could also be explored to observe if jumping is a factor in shot making.

It is also recommended that when performing studies related to machine learning, time series and skeletal data, gather as much data as possible for training and testing machine learning models. This allows the machine learning models to perform well and be more effective and accurate. When performing training and testing the model, it is important that the data to be used for both sets should be properly balanced.

The placement and angle of the camera should also be considered since there are other certain unique features that could

be gathered from other camera angles and camera positions that are not tested in this study that could be helpful for the model in classifying shots. The shot trajectory of both shooting forms could also be considered as a feature to be included in the machine learning models because the study focuses on the skeletal points of the shooting form as the sole basis of the classifications, wherein the shot trajectory could also be a factor in classifying the shot.

REFERENCES

[1] G. Büyükbaykal, "Importance of sports journalism education," *Journalism and Mass Communication*, vol. 6, no. 11, pp. 661–668, 2016.

[2] J. F. Drazan, A. K. Loya, B. D. Horne, and R. Eglash, "From sports to science: Using basketball analytics to broaden the appeal of math and science among youth," *In Conference Paper presented at Rensselear Polytechnic Institute, NY*, 2017. [Online]. Available: https://www. researchgate.net/publication/314263728

[3] A. Taniguchi, K. Watanabe, and Y. Kurihara, "Measurement and analyze of jump shoot motion in basketball using a 3-d acceleration and gyroscopic sensor," *In 2012 Proceedings of SICE Annual Conference (SICE)*, pp. 361–365, Aug. 2012.

[4] A. Hölzemann and K. Van Laerhoven, "Using wrist-worn activity recognition for basketball game analysis," *In Proceedings of the 5th international Workshop on Sensor-based Activity Recognition and Interaction*, p. 13, Sep. 2018.

[5] J. C. Maglott, J. Xu, and P. B. Shull, "Differences in arm motion timing characteristics for basketball free throw and jump shooting via a body-worn sensorized sleeve," *In 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 31–34, May 2017.

[6] M. Nakai, Y. Tsunoda, H. Hayashi, and H. Murakoshi, "Prediction of basketball free throw shooting by openpose," *In Proceedings of the Fifth International Workshop on Skill Science, SKL*, vol. 18, 2018.

[7] Y. Khan, N. Khan, S. Farooq, A. Abid, S. Khan, F. Ahmad, and M. Mahmood, "An efficient algorithm for recognition of human actions," *The Scientific World Journal*, vol. 2014, pp. 2562–2572, 2014.

[8] Y. Gu, H. Do, Y. Ou, and W. Sheng, "Human gesture recognition through a kinect sensor," *In 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1379–1384, Dec. 2012.

Exploring the Persuasiveness of An Emotional Intelligence Training Application using U-FADE

Alyssa P. Jubay, Yna Mari D. Ojeda, Ma. Rowena C. Solamo, Rommel P. Feria, Ligaya Leah Figueroa
Department of Computer Science
University of the Philippines, Diliman
{apjubay|ydojeda|rpferia|rcsolamo|llfigueroa}@up.edu.ph

ABSTRACT

The research explores the persuasiveness of an emotional intelligence training (EIT) mobile application, called *Maintein*, using the Unified Framework for Analyzing, Designing and Evaluating Persuasive Systems (U-FADE), which asserts the need to have a before-and-after analysis of attitude and behavior during the period of using the application. Theory of Planned Behavior (TPB) is used for this analysis. An experiment has been conducted with twenty-five (25) student volunteers who are tasked to use *Maintein* for a period of seven (7) weeks. The TPB Preliminary Questionnaire is answered before the experiment. The TPB Post Questionnaire and Persuasiveness of U-FADE System Features Questionnaire are answered after the experiment. *Maintein* supports reflective journaling, active listening and goal tracking as the EIT exercises, and the monitoring of emotional and mental well-being states. To inform its persuasive design, the study reveals (1) that messages using past behavior may be used to encourage initial and continuous performance of the EIT activities, and the monitoring of emotional and mental well-being, (2) that the Active Listening Module needs to be reanalyzed and redesigned, (3) that simple interactions that lead to its ease of use should guide implementation, (4) that tailored content is needed for prompts in reflective journaling and active listening, and (5) that further analysis on the frequency, timing and message framing of reminders and praise needs to be conducted so that they can be used not to impede but rather to encourage performance of the EIT exercises and monitor emotional and mental well-being states.

KEYWORDS

Mental Well-being, Emotional Intelligence, Emotional Intelligence Training, Persuasive Technology, U-FADE

ACM Reference Format:
Alyssa P. Jubay, Yna Mari D. Ojeda, Ma. Rowena C. Solamo, Rommel P. Feria, Ligaya Leah Figueroa. 2020. Exploring the Persuasiveness of An Emotional Intelligence Training Application using U-FADE. In *Proceedings of (UP Diliman).* UP Diliman, Quezon City, Philippines, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

UP Diliman, May 2020, Quezon City, Philippines

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Emotional Intelligence (EI) is one of the newly coined terms in Psychology. It is defined as *"the ability to perceive, manage, and assess your emotions, and others' emotions"* [13]. It is commonly known as the emotional quotient of an individual where it describes his or her level of emotional intelligence which is often represented by a score. According to Serrat, there are five domains of emotional intelligence: self-awareness, self-regulation, self-motivation for personal competencies, social awareness and social skills for social competencies [13]. Many studies have examined the predictive nature of EI and its domain to mental health. Students with high EI have been revealed to be less likely to externalize problems and be emotionally adjusted, than students with lower EI who are found to be prone to engage in harmful behaviors [3][4]. Because EI has been found to moderate the association between stressful experiences and suicidal behaviors, it is also a protective factor in at-risk adolescent populations [5].

The emotional intelligence of an individual can increase by undergoing several emotional intelligence training (EIT) or interventions. Studies show that individuals who undergo emotional intelligence training sessions have their emotional intelligence increased than those individuals who do not. For example, in Bagheri et al. study, after six sessions of emotional intelligence training that includes emotion perception, understanding emotions, using emotions to facilitate thought and managing emotions, the experiment shows that undergraduate students who undergo these training sessions have their emotional intelligence increased than those who do not [2]. Based on the same study, the training has improved one's mental health, forms better social relationships, and adapts to a stressful environment or event better [2]. Another study examined the short and midterm effects of emotional intelligence training on adolescent health. Students who participated in the EIT of the study reported lower negative affect scores and lower number of clinical symptoms, including anxiety, social stress, depression, external locus of control, sense of incapacity, and somatization [11]. The students also reported an increase MH-5 measure of mental health. Their results persisted for at least 6 months after intervention.

Persuasive technology is a tool that automates the increase in a person's capacity for behavioral change [6]. It is designed to provide an experience that may motivate reflection or rehearsal, and to provide response or guidance to user interaction [9]. To help in the design, implementation and evaluation of persuasive systems, the Unified Framework for Analyzing, Designing and Evaluating Persuasive Systems (U-FADE) can be employed. It is a framework that provides a systematic approach that facilitates persuasive design, and addresses the issue of changing needs of users by considering external and internal activities that can possibly promote or impede

persuasion before and after implementation. U-FADE compiled 28 system features that support peripherally or elaborated routed messages and that serves as a guide in selecting system features during persuasive design. These 28 features, which can be found in Table 1 were categorized according to the type of message they readily prompted [14].

Primary Tasks Support	Dialogue Support	System Credibility	Social Support
Tunneling	Suggestion	Expertise	Social Learning
Self-monitoring	Social Role	Verifiability	Social Comparison
Simulation	Rewards	Authority	Social Learning
Rehearsal	Reminder	Surface Credibility	Recognition
Personalisation	Similarity	Trustworthiness	Competition
Reduction	Linking	Real-World Feel	Cooperation
Tailoring	Praise	Authority	Normative Influence

Table 1: U-FADE System Features supporting either elaborated or peripheral routes

U-FADE asserts the necessity to evaluate actual changes in attitude and behavior that occur during the use of the persuasive system. One model that can be used is the 3-Dimensional Relationship between Attitude and Behavior (3D-RAB) Model. There are three (3) internal factors to consider for analyzing the state of a user (current behavior, attitude towards target behavior, attitude towards changing and maintaining the current behavior), and two (2) external factors (natural attitude or behavioral change and planned behavior change.)

Another model that can be used to evaluate changes in attitude and behavior is Theory of Planned Behavior (TPB). It is used to understand and predict behaviors [8]. It is similarly guided by three considerations: *behavioral beliefs* which are the beliefs about the likely consequences and experiences associated with a behavior, *normative beliefs* which are beliefs about the normative expectations and behaviors of others, and *control beliefs* which are beliefs about the presences of factors that may facilitate or impede performance of behavior, and produce *perceived behavioral control* [1]. According to TPB, the more favorable the attitude and subjective norm, the stronger the intention to perform the target behavior. If given a sufficient degree of actual control over the behavior, it can be expected for intentions to be carried out when an opportunity arises. It views perceived behavioral control as a proxy for actual control, and intention is assumed to be the immediate antecedent of behavior.

Having a before-and-after analysis allows the identification of other related changes in the user which is not explicit, and yet impacts the target behavior. For example, after using an application to promote eating vegetables, a user may change his attitude towards vegetables and yet may not be eating them. Such information can inform the developers on what to do next since the intent of the design is not achieved. [15].

The objective of the research is to examine the persuasiveness of Maintein, an emotional intelligence training mobile application that focuses on encouraging users to regularly exercise their emotional intelligence and monitor their emotional and mental states. It uses Unified Approach to Persuasive Systems Development Framework (U-FADE) to design, implement and evaluate. The Theory of Planned Behavior (TPB) is used to understand the changes in the attitude and behavior of the participants who will use the system.

The study is limited to the undergraduate students. The research experiment covered the weeks between March 10 to April 27, 2020. Due to the 2019-nCov Acute Respiratory Disease (2019-nCov ARD) pandemic, classes were suspended at all levels. During these times, class gatherings, including online classes, were highly discouraged. Because the application was designed for college students and college life, the circumstances brought about by the pandemic may have affected the participants' mental well-being and their participation in the experiment.

2 MAINTEIN: SOFTWARE ARCHITECTURE

Maintein is a persuasive application that focuses on encouraging users to regularly exercise their emotional intelligence. By increasing their emotional intelligence, users are better equipped in adapting to stressful environments and maintaining their positive mental well-being. There are three emotional intelligence training (EIT) activities that were supported in the application, namely, *Reflective Journal* for developing self-awareness and self-regulation, *Active Listening* for social awareness and social skills, and *Goal Tracking* for developing self-motivation. A Mental Health Monitoring module serves as an assessment that facilitates monitoring of emotional intelligence (EI) and mental well-being (MWB) states.

2.1 Reflective Journaling Module

Reflective Journaling (RJ) helps develop self-awareness and self-regulation. Both entail looking inwards to oneself. The module facilitates reflective journal writing. A key feature of the module is the reflection prompts that help users start the reflection writing by giving out starting phrases. This can be seen in Figure 1.

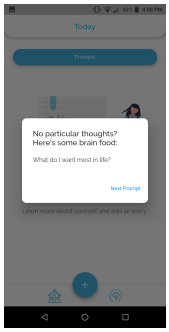


Figure 1: Reflection Prompts of Reflective Journal

2.2 Active Listening Module

Active Listening (AL) demands a person to place their undivided attention and awareness at the speaker's disposal, to listen to the speaker with interest and to appreciate without interrupting[10]. It involves verbal and non-verbal communications. It is sometimes defined more of what is not done by what is. It helps develop social awareness and social skills. The module facilitates active listening. It has a listener checker that guides users in paraphrasing conversations, and making users more conscious of traits or characteristics

of active listening that they portrayed through a checklist. A key feature of the module is the conversation prompts that is helpful to start conversations with others. This can be seen in Figure 2.

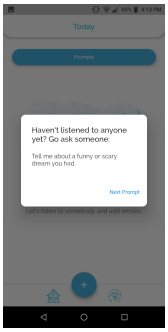


Figure 2: Conversation prompts of Active Listening

2.3 Goal Tracking Module

Goal setting and tracking helps develop self-motivation. It is used as an EIT activity where emotions motivate certain situation. For instance, Gill described the use of goal setting in basketball season [7]. According to him, goal setting is one of the common and popular strategies used as it facilitates direction and focus to raise their performance level. The module allows the user to set goals with deadlines, levels of urgency, subtasks, and expected outputs. Users are alerted of the deadlines by push notifications, and may mark each goal, output, and subtask complete or incomplete. Users may create such goals, view them, and edit each goals' details. Figure 3 shows the goal form, goal list and detail screen of this module.

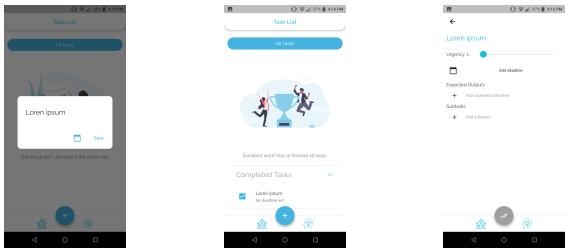


Figure 3: Goal Form, Goal List, Detail Screen of Goal Tracking

2.4 Mental Health Monitoring Module

The Mental Health Monitoring Module facilitates monitoring of the user's mental well-being (MWB) and emotional intelligence (EI) scores. The Emotional Intelligence Scale (EIS) was used in Maintein. It draws on the ability of the model, conceptualizing emotional intelligence in terms of potential for intellectual and emotional growth [12]. Thus, it assesses the ability to process information about one's own and other's emotions. The scale is free for non-commercial use. Each question was presented in a single page within a screen. This is shown in Figure 4.

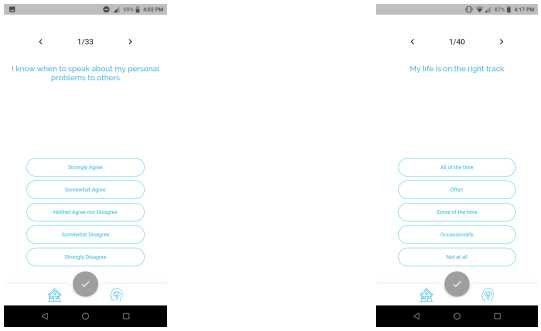


Figure 4: EIS and MWB Assessment Forms

2.5 Settings & Notification Module

The Settings Screen which holds the configurations for push notifications, information about the application, appropriate attributions, and the participant identification number can be seen in Figure 5.

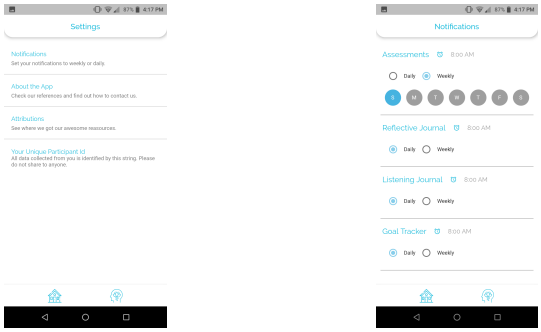


Figure 5: Settings & Notifications

The development of mobile application was done using Flutter, Google's open-source UI software development kit for building natively compiled applications from a single codebase. Flutter uses the Dart programming language and allows support for both Android and iOS devices. Flutter was chosen for its dynamic and fast cross-platform development.

2.6 U-FADE System Features in Maintein

U-FADE provides a list of possible system features for persuasive design as seen in Table 1. In this study, five (5) system features were incorporated in Maintein and explored their persuasiveness. Table 2.6 lists them.

U-FADE System Feature	Maintein Design and Implementation
Self-Monitoring	Maintein offers assessments that keep track of the EI and MWB scores.
Reduction	Activities in Maintein are broken into steps. Entry inputs consist of multiple screens to reduce data input at a given time.
Suggestion	Reflection prompts and conversation prompts are available for reflective journaling and active listening.
Reminder	Maintein allows users to set daily or weekly notification for each of the activities it facilitates.
Praise	Maintein provides positive feedback when users do activities in the application.

Table 2.6: U-FADE System Features in Maintein

3 RESEARCH METHODOLOGY

The study conducted an experiment of using Maintein where participants are asked to download the application and use it over a 7-week period. To evaluate changes in attitudes and behavior, a preliminary questionnaire and a post-questionnaire for the before-and-after analysis were developed and used. Another questionnaire was developed and used after the experiment to understand the persuasiveness nature of the U-FADE System Features in Maintein.

3.1 Participants

There were a total of 39 volunteers who consented to participate in the study. All participants were undergraduate students. All volunteers were briefed about the research and the extent of their participation. Afterwards, they signed the consent form. All participation was completely voluntary, and participants were free to quit at any point of the experiment. The experiment started on March 10, 2020. This coincide with the suspensions of classes due to the 2019 nCOV ARD pandemic. Because of this, only 31 of the 39 participants answered the preliminary survey, and proceeded with the experiment. By the last day of the experiment, only 25 of the participants opted to finish the experiment and all the surveys. The data collected from the 14 volunteers who did not complete their participation were removed from the data set. Only the data from the 25 participants who completed the experiment and all the surveys they answered were used for the analysis of the study.

3.2 Research Instruments

To evaluate the actual changes in attitude and behavior that occur during the experiment, Theory of Planned Behavior (TPB) was used to guide the development of the *TPB Preliminary Questionnaire* and *TPB Post Questionnaire*. Factors that were considered are past behavior, intention, perceived behavioral control, subjective norm and attitude focusing on EIT activities of reflective journaling, active listening, goal tracking and EI and MWB monitoring. *Persuasiveness of U-FADE System Features Questionnaire* was also developed to explore the persuasiveness of U-FADE System Features in Maintein (as seen in Table 1).

3.2.1 TPB Preliminary Questionnaire. The TPB Preliminary Questionnaire is used for the before analysis on the attitudes and behavior of the participant. Table 3.2.1 shows the factors and corresponding questions with the scale used.

TPB Factors	Statement	Scale Used
Past Behavior	Choose the longest duration you've gone doing the activities or similar activities to the ones below. [Reflective Journaling]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	Choose the longest duration you've gone doing the activities or similar activities to the ones below. [Active Listening]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	Choose the longest duration you've gone doing the activities or similar activities to the ones below. [Goal Tracking/Goal Setting]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	Choose the longest duration you've gone doing the activities or similar activities to the ones below. [Mental Health Tracking/Monitoring]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	Choose the longest duration you have gone using applications that facilitate activities or similar activities to the ones below. [Reflective Journaling]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	Choose the longest duration you have gone using applications that facilitate activities or similar activities to the ones below. [Active Listening]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	Choose the longest duration you have gone using applications that facilitate activities or similar activities to the ones below. [Goal Tracking/Goal Setting]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	Choose the longest duration you have gone using applications that facilitate activities or similar activities to the ones below. [Mental Health Tracking/Monitoring]	5-point scale: Never, Days, Weeks, Months, Years
Intention	Check the intensity level of your intention to do the following activities using the application. [Reflective Journaling]	5-point scale: 1 - little to no intensity and intention, 5 - highest intensity and intention.
Intention	Check the intensity level of your intention to do the following activities using the application. [Active Listening]	5-point scale: 1 - little to no intensity and intention, 5 - highest intensity and intention.
Intention	Check the intensity level of your intention to do the following activities using the application. [Goal Tracking/Goal Setting]	5-point scale: 1 - little to no intensity and intention, 5 - highest intensity and intention.
Intention	I intend to do the activities I've stated for the following duration.	Number of months
Perceived behavioral control	I am confident that I can do my selected activities for the duration I've stated.	5-point scale: 1 - not confident, 5 - very confident
Perceived behavioral control	My doing of the activities I've selected for the stated duration is solely up to me.	5-point scale: 1- disagree, 5 - agree
Subjective norm	Most people who are important to me would approve of me doing those activities.	5-point scale: 1- disagree, 5 - agree
Subjective norm	Most people I know would do those activities or similar activities.	5-point scale: 1- disagree, 5 - agree
Attitude	Doing the activities I've chosen would be...	5 point scale:1 - bad, 5 - good.

Table 3.2.1: Theory of Planned Behavior (TPB) Preliminary Questionnaire

3.2.2 TPB Post Questionnaire. The TPB Post Questionnaire is used for the after analysis of the attitude and behavior of the participants. It was generated from the statements of the TPB Preliminary Questionnaire. Table 3.2.2 shows the factors and corresponding questions with the scale used.

TPB Factors	Statement	Scale Used
Past Behavior	After completing the experiment, choose the longest duration you've gone doing the activities using Maintein. [Reflective Journaling]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	After completing the experiment, choose the longest duration you've gone doing the activities using Maintein. [[Active Listening]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	After completing the experiment, choose the longest duration you've gone doing the activities using Maintein. [Goal Tracking/Goal Setting]	5-point scale: Never, Days, Weeks, Months, Years
Past Behavior	After completing the experiment, choose the longest duration you've gone doing the activities using Maintein. [Mental Health Tracking/Monitoring]	5-point scale: Never, Days, Weeks, Months, Years
Intention	After completing the experiment, check the intensity level of your intention to continue doing the following activities using the application. [Reflective Journaling]	5-point scale: 1 - little to no intensity and intention, 5 - highest intensity and intention.
Intention	After completing the experiment, check the intensity level of your intention to continue doing the following activities using the application. [Active Listening]	5-point scale: 1 - little to no intensity and intention, 5 - highest intensity and intention.
Intention	After completing the experiment, check the intensity level of your intention to continue doing the following activities using the application. [Goal Tracking/Goal Setting]	5-point scale: 1 - little to no intensity and intention, 5 - highest intensity and intention.
Intention	After completing the experiment, I intend to continue doing the activities I've stated for the following duration.	Number of months
Perceived behavioral control	I am confident that I can do my selected activities for the duration I've stated.	5-point scale: 1 - not confident, 5 - very confident
Perceived behavioral control	My usage of Maintein for the past weeks was solely up to me.	5-point scale: 1- disagree, 5 - agree
Subjective norm	Now that I've used the application, I find that most people who are important to me would approve of me using Maintein.	5-point scale: 1- disagree, 5 - agree
Subjective norm	Most people I know would use Maintein or similar applications.	5-point scale: 1- disagree, 5 - agree
Attitude	After using Maintein, I can say that doing the activities in it would be…	5 point scale:1 - bad, 5 - good.

Table 3.2.2: Theory of Planned Behavior Post Questionnaire

3.2.3 Persuasiveness of U-FADE System Features Questionnaire. The Persuasiveness of U-FADE System Features Questionnaire is used to explore the persuasiveness of U-FADE System Features in Maintein as shown in Table 1. Table 3.2.3 shows the U-FADE System Feature and corresponding questions with the scale used.

U-FADE System Feature	Statement	Scale Used
Self-monitoring	The assessments module made monitoring my mental well-being and emotional intelligence easier.	5-point Scale: 1 - Strongly Disagree, 5 - Strongly Agree
Reduction	The activities in the application are reduced to steps that helped me perform them smoothly	5-point Scale: 1 - Strongly Disagree, 5 - Strongly Agree
Suggestion	I often use the reflection prompts to make an entry.	5-point Scale: 1 - Strongly Disagree, 5 - Strongly Agree
Suggestion	I often use the conversation prompts to make an entry.	5-point Scale: 1 - Strongly Disagree, 5 - Strongly Agree
Praise	Praises and feedback from using the application encouraged me to do activities regularly.	5-point Scale: 1 - Strongly Disagree, 5 - Strongly Agree

Table 3.2.3: Persuasiveness of U-FADE System Features Questionnaire

3.3 Data Collection

Each participant had a unique Participant ID (PID) that could be found on the application settings. The application installer was accessed by the participants through an online link. Once they had access to their PID, the TPB Preliminary questionnaire on attitudes and behavior was answered online. Afterwards, the experiment started and participants are asked to use the application for seven (7) weeks. After that, participants answered the two (2) post-questionnaires- one on attitude and behavior, and the other on persuasiveness of U-FADE System Features online. Only the data of the 25 participants who finished the experiment and answered all surveys were kept, and their PIDs were replaced with identifiers ID1 to ID25.

4 ANALYSIS AND RESULTS

4.1 Changes in Attitude and Behavior using Theory of Planned Behavior

U-FADE insists of having a before-and-after analysis in attitude and behavior during the course of using a persuasive system. The results of which are used to guide improvements on the persuasive design of the application. The factors that are considered in the study are *past behavior*, *intention*, *perceived behavioral control*, *subjective norms* and *attitude*.

4.1.1 Past Behavior. Questions on past behavior were used to find out participants' experience and familiarity of doing the three (3) emotional intelligence exercises (i.e., Reflective Journaling, Active Listening and Goal Tracking), and monitoring their mental health. They were also used to uncover if doing activities with Maintein has something to do with their past behavior. A Chi-square Test of Independence was used to determine whether the behavior of the participants on doing the activities after the experiment is associated with the behavior they had before doing the experiment.

Reflective Journaling. Table 4.1.1a shows the comparison of the data collected for doing Reflective Journaling before and after the experiment. Prior to the experiment, most respondents have experience in reflective journaling. 24% of the respondents have experience doing it for days, 8% of the respondents have experienced it doing for weeks, and 28% have experienced doing it for months. However, not all respondents are familiar with mobile applications that facilitate reflective journaling. 68% of those who answered "Never" have not tried mobile applications that facilitate reflective journaling while 32% have experienced mobile applications that facilitate reflective journaling. After the experiment, the majority of the respondents have tried using reflective journaling on Maintein for days. The table shows that there was an increase in the number of users who experience doing reflective journaling. It may suggest that the participants were persuaded to try reflective journaling or were curious about it. A chi-square test was performed on Table 4.1.1a on past behavior of reflective journaling before the experiment (first column) and past behavior of reflective journaling after the experiment (third column). The difference between the two is significant, X^2 (3, N=25) = 14.13, p = 0.002. This suggests that reflective journaling will most likely be used by those who have experienced it before than those who do not.

Duration	**Choose the longest duration you've gone doing the activities or similar activities to the ones below.[Before] Number of Respondents / Percentage**	**Choose the longest duration you have gone using applications that facilitate activities or similar activities to the ones below.[Before] Number of Respondents / Percentage**	After completing the experiment, choose the longest duration you've gone doing the activities using Maintein. Number of Respondents / Percentage
Never	10 / 40%	17 / 68%	5/ 20%
Days	6 / 24%	8 / 32%	17 / 68%
Weeks	2 / 8%	0 / 0%	3 / 12%
Months	7 / 28%	0 / 0%	0 / 0%

Table 4.1.1a: Results of Reflective Journal on Past Behavior

Active Listening. Table 4.1.1b shows the tabulation on past behavior of active listening. Prior to the experiment, more than half of the respondents are not familiar with active listening exercises. 8% each of the respondents have tried active listening exercises for days and weeks while 28% of the total respondents tried it for months. The majority of the participants have not tried using mobile applications that facilitate active listening. 8% of them have tried using mobile applications that facilitate active listening for days while 4% of them have tried using it for months. After the experiment, the majority still have not experience doing active listening on Maintein. This may be attributed to participants might have different expectations of what active listening is as there are many different types that was not considered in the study. Also, they have different initial experience than what the module offers. A chi-square test was performed on Table 4.1.1b on past behavior of active listening before the experiment (first column) and past behavior of active listening after the experiment (third column). The difference between the two is significant, X^2 (3, N=25) = 8.4242, p = 0.03801 which suggests that active listening will most likely be used by those who used it before than those who do not have experience with it.

Duration	**Choose the longest duration you've gone doing the activities or similar activities to the ones below.[Before] Number of Respondents / Percentage**	**Choose the longest duration you have gone using applications that facilitate activities or similar activities to the ones below.[Before] Number of Respondents / Percentage**	After completing the experiment, choose the longest duration you've gone doing the activities using Maintein. Number of Respondents / Percentage
Never	14 / 56%	22 / 88%	19/ 76%
Days	2 / 8%	2 / 8%	4 / 16%
Weeks	2 / 8%	0 / 0%	2 / 8%
Months	7 / 28%	1 / 4%	0 / 0%

Table 4.1.1b: Results of Active Listening on Past Behavior

Goal Tracking. Table 4.1.1c shows the comparison of the data collected for Goal Tracking before and after the experiment. Before the experiment, most of the respondents have experienced goal tracking. 24% of the respondents have never experienced goal tracking, 28% of the respondents have experienced it doing for days, while 24% each of the total respondents have experienced it for weeks and months. Most respondents have experienced using mobile applications with goal tracking on it. Only 32% of the respondents have never tried using goal tracking mobile applications. After the experiment, 56% of the total respondents have tried using the goal tracking module on Maintein, 20% of the total respondents have tried it for weeks, and only 24% have never used them. The table shows that there was an increase in the number of users who experience doing goal tracking. It may suggest that the participants were persuaded to put and define their goals. A chi-square test was performed on Table 4.1.1c on past behavior of goal tracking before the experiment (first column) and past behavior of goal tracking after the experiment (third column). The difference between the two is significant, X^2 (3, N=25) = 8.4242, p = 0.03801, which suggests that goal tracking will most likely be used by those who like to put and define their goals than those who do not.

Duration	**Choose the longest duration you've gone doing the activities or similar activities to the ones below.[Before] Number of Respondents / Percentage**	**Choose the longest duration you have gone using applications that facilitate activities or similar activities to the ones below.[Before] Number of Respondents / Percentage**	After completing the experiment, choose the longest duration you've gone doing the activities using Maintein. Number of Respondents / Percentage
Never	6 / 24%	8 / 32%	6/ 24%
Days	7 / 28%	10 / 40%	14 / 56%
Weeks	6 / 24%	3 / 12%	5 / 20%
Months	6 / 24%	4 / 16%	0 / 0%

Table 4.1.1c: Results of Goal Tracking on Past Behavior

Mental Health Monitoring. Table 4.1.1d shows the comparison of the data collected for monitoring of mental health before and after the experiment. Prior to the experiment, only 40% of the total respondents have not tried Mental Health Monitoring. The data on the column for mobile applications showed that more than half of the participants have not experienced using mobile applications that facilitate mental health monitoring. After the experiment, it shows that only 8% of the respondents have never used the mental health monitoring module on Maintein. 10% each of the respondents has experienced using the mental health monitoring module on Maintein while 12% of the respondents have used it for months. The increase in the number of users who use mental health monitoring suggests that they monitor their mental health during the period of the experiment. A chi-square test was performed on Table 4.1.1d on past behavior of mental health monitoring before the experiment (first column) and past behavior of mental health monitoring after the experiment (third column). The difference between the two is significant, X^2 (3, N=25) = 10.9143, p = 0.0122, which suggests that mental health monitoring will most likely be used by those who monitor their mental well-being than those who do not.

Duration	Choose the longest duration you've gone doing the activities or similar activities to the ones below.[Before] Number of Respondents / Percentage	Choose the longest duration you have gone using applications that facilitate activities or similar activities to the ones below.[Before] Number of Respondents / Percentage	After completing the experiment, choose the longest duration you've gone doing the activities using Maintein. Number of Respondents / Percentage
Never	10 / 40%	14 / 56%	2 / 8%
Days	11 / 44%	8 / 32%	10 / 40%
Weeks	2 / 8%	2 / 8%	10 / 40%
Months	2 / 8%	1 / 4%	3 / 12%

Table 4.1.1d: Results of Mental Health Monitoring on Past Behavior

In general, what the results on past behavior show is that the use of Maintein in doing emotional intelligence training of reflective journaling, active listening, goal tracking, and monitoring of mental health will most likely be done by those who do doing emotional intelligence training of reflective journaling, active listening, goal tracking, and monitoring of mental health. This informs the persuasive design to use past behavior as part of the messaging to encourage initial use of the application or to encourage maintaining the behavior (of reflective journaling, active listening, goal tracking and monitoring of mental health) with the Maintein. Message themes such as *"You were able to do it before, you can do it now."* may help.

4.1.2 Intention. The items in Table 4.1.2 shows the average scores for intention. It discusses the intensity of the intention of the users in doing the emotional intelligence training, i.e., reflective journaling, active listening and goal tracking.

Statement	Preliminary Experiment Average Response ± Standard Deviation	Post Experiment Average Response ± Standard Deviation
Check the intensity level of your intention to do the following activities using the application. [Reflective Journaling]	2.88 ± 1.13	2.92 ± 1.35
Check the intensity level of your intention to do the following activities using the application. [Active Listening]	2.52 ± 1.16	1.72 ± 1.02
Check the intensity level of your intention to do the following activities using the application. [Goal Tracking/Goal Setting]	3.16 ± 1.28	2.92 ± 1.41

Table 4.1.2: Results of Intention on the Emotional Intelligence Training Activities

Reflective Journal. Prior to the experiment, the average intensity of the intention of the participants is 2.88. After experiencing reflective journaling on Maintein during the 7-week experiment, the average intensity level of the intention of the participants is 2.92. Even if there was a slight increase from the preliminary experiment, it shows that the intensity level of the user remains neutral. As per the results of a paired t-test, there is no significant difference in the scores of the intention in doing reflective journaling before the experiment (M=2.88, SD=1.13) and after the experiment (M=2.92, SD=1.35); t(24) = -0.14, p = 0.88.

Active Listening. Prior to the experiment, the average intensity of the intention of the participants is 2.52. After 7 weeks, the average intensity of the intention of the participants is 1.72. It shows that from a level 2 intensity, it became a level 1 intensity. After performing a paired t-test, there was a significant difference in the scores for the intensity level on the user's intention before the experiment (M = 2.52, SD =1.16) and after the experiment (M = 1.72, SD = 1.02); t(24) = 2.06, p = 0.001. The difference in the intensity of the intention in doing active listening may be attributed to several factors. As mentioned in the previous section, participants might have different expectations of what active listening is as there are many different types that was not considered in the study. They have different initial experience than what the module offers. Another, because of the COVID-19 Pandemic which forced people to stay at home, participants have limited option to whom they can talk to as active listening requires a more social circumstances.

Goal Tracking. Prior to the experiment, the average intensity of the intention of the participants to do goal tracking is 3.16. After the 7-week experiment, the average intensity level of the intention of the users slightly decreased to 2.92. Even if there was a decrease in the average intensity of the intention of the participants, it shows that the intensity level of the intention of the users remains neutral. As per the results of a paired t-test, there is no significant difference in the scores of the intention in doing goal tracking before the experiment (M=3.16, SD=1.28) and after the experiment (M=2.92, SD=1.41); t(24) = 1.24, p = 0.22.

What the results of intention show, in terms of informing the persuasive design, is that Active Listening Module needs to be reanalyzed and redesigned to address the different expectations of the participants, probably, consider an active listening exercise that is commonly used or popular.

4.1.3 Perceived Behavioral Control. Table 4.1.3 shows the results of the perceived behavioral control, which are the beliefs of the user about the presence of factors that may facilitate or impede the performance of selected EIT activities and monitoring of mental health.

Statement	Preliminary Experiment Average Response ± Standard Deviation	Post Experiment Average Response ± Standard Deviation
I am confident that I can do my selected activities for the duration I've stated.	3.6 ± 1.04	3.12 ± 0.93
My usage of Maintein for the past weeks was solely up to me.	4.4 ± 0.65	4.32 ± 0.80

Table 4.1.3: Results of Perceived Behavioral Control on Emotional Intelligence Training Activities and Monitoring of Mental Health

The average score on the confidence of the participants to do the selected activities prior to the experiment is 3.6. After the 7-week experiment, the average confidence is 3.12. Nothing has changed in terms of the level of confidence of the participants, and it remains neutral.

The average score on the self-efficacy of the participants in doing the selected activities prior to the experiment 4.4. After they have done experiment, the average score of self-efficacy is 4.32. Even though there was a slight decrease between the two samples,

nothing has changed in terms of the level of self-efficacy of the participants, and it remains relatively high (i.e., they can do them).

The results of perceived behavioral control doesn't provide any guidance to the improvements on the persuasive design of Maintein.

4.1.4 Subjective Norm. Table 4.1.4 shows the results on subject norms, which are beliefs about the normative expectations and behaviors of the significant others.

Statement	Preliminary Experiment Average Response ± Standard Deviation	Post Experiment Average Response ± Standard Deviation
Now that I've used the application, I find that most people who are important to me would approve of me using Maintein.	4.44 ± 0.71	4.12 ± 0.73
Most people I know would use Maintein or similar applications.	3.44 ± 1.00	3.12 ± 1.20

Table 4.1.4: Results of Subjective Norms on Emotional Intelligence Training Activities and Monitoring of Mental Health

Prior to the experiment, the score on the participants belief on others approving their use of Maintein is 4.44. After the experiment, it became 4.12. Nothing has changed, in terms of this belief, and it remains relatively high. Similarly, the score on the belief that others will use Maintein (or similar application) before the experiment is 3.44. After the experiment, it became 3.12. Nothing has changed, in terms of this belief, and it remain neutral.

The results on subjective norms doesn't provide any guidance on the improvements on the persuasive design of Maintein.

4.1.5 Attitude. Table 4.1.5 shows the results of the attitude of the individual in performing emotional intelligence training activities and monitoring mental health.

Statement	Preliminary Experiment Average Response ± Standard Deviation	Post Experiment Average Response ± Standard Deviation
I can say that doing the activities in it would be… [5-good or 1-bad]	4.6 ± 0.5	4.6 ± 0.58

Table 4.1.5: Results of Attitude on Emotional Intelligence Training Activities and Monitoring of Mental Health

According to the participants, doing the activities is good for them for both before and after the experiment. There was no difference in the average score (4.6) on attitude before and after the experiment. It remains good, ie., they have favorable attitude towards doing the activities. The results on attitude doesn't provide any guidance on the improvements on the persuasive design of Maintein.

In summary, analyzing past behavior found that that the use of Maintein in doing emotional intelligence training of reflective journaling, active listening, goal tracking, and monitoring of mental health will most likely be done by those who do doing emotional intelligence training of reflective journaling, active listening, goal tracking, and monitoring of mental health. Past behavior may encourage them to perform the activities in the future. This informs the persuasive design to use past behavior as part of the messaging to encourage initial use of the application or to encourage maintaining the behavior (of reflective journaling, active listening, goal tracking and monitoring of mental health) with the application.

Analyzing intention, on the other hand, showed that the Active Listening module needs to be reanalyze and redesign, in terms, of the expectations of the module. The results for perceived behavioral control, intention, subjective norms and attitude did not provide any guidance in the improvement of the persuasive design of Maintein.

4.2 Persuasiveness of U-FADE System Features in Maintein

Table 4.2 shows the results of the Persuasiveness of U-FADE Systems Features in Maintein.

U-FADE System Feature	Statement	Average Response ± Standard Deviation
Self-monitoring	The assessments module made monitoring my mental well-being and emotional intelligence easier. [5-Strongly agree or 1-Strongly disagree]	4.12 ± 1.20
Reduction	The activities in the application are reduced into simple steps that helped me perform them smoothly. [5-Strongly agree or 1-Strongly disagree]	4.24 ± 0.60
Suggestion	I often use the reflection prompts to make an entry [Reflection Journal]. [5-Strongly agree or 1-Strongly disagree]	2.56 ± 1.12
Suggestion	I often use the conversation prompts to start conversations to listen to. [Active Listening] [5-Strongly agree or 1-Strongly disagree]	2.2 ± 1.26
Reminder	The app's notifications persuaded me to use the application when notified. [5-Strongly agree or 1-Strongly disagree]	2.92 ± 1.04
Praise	Praise and feedback from using the application encourages me to do the activities regularly. [5-Strongly agree or 1-Strongly disagree]	3.44 ± 1.00

Table 4.2: Results of Persuasiveness of U-FADE Features in Maintein

Self-monitoring through the Mental Health Tracking Module of emotional intelligence and mental well-being scores of the participants and *Reduction* through simple steps yielded a relative high score (4.12 and 4.24 respectively) in ease of use. This indicates, in terms of persuasive design, to continue having simple interactions that makes Maintein easy to use.

On the other hand, *suggestion* through reflection and conversation prompts, *reminders through push notifications* and *praise* through feedback have low to moderate scores (2.56 and 2.2, 2.92 and 3.44 respectively). For *Suggestion*, the reflection prompts through further investigation show that participants look to reflection prompt but the prompt fails to help them start the reflection. This suggests that content tailored for easier reflections may be suited for the prompts. The conversation prompts through further investigation reveal that participants did not use the prompts precisely because of the issue that arose with using Active Listening Module as discussed in the previous sections. For *Reminder* and *Praise*, further investigation show that factors that affected the persuasiveness of the features are frequency, timing and message framing.

5 CONCLUSION

The study was about exploring the persuasiveness of Maintein, an emotional intelligence training mobile application using the Unified Framework of Analyzing, Designing and Evaluating Persuasive Systems (U-FADE). An experiment was done involving twenty-five (25) student volunteers, in which, they used the application for a seven

week period. They answered three (3) questionnaires- TPB Preliminary Questionnaire, TPB Post Questionnaire and Persuasiveness of U-FADE System Features Questionnaire. The before-and-after analysis on the attitude and behavior revealed that:

(1) the use of past behavior needs to be part of the messaging to encourage initial use or maintaining the behavior of doing emotional intelligence exercises and monitoring of mental health;
(2) the module on Active Listening needs to be reanalyze and redesign;
(3) perceived behavioral control, intention, subjective norms and attitude did not yield much result to inform the persuasive design of Maintein.

On the other hand, the analysis of the persuasiveness of the U-FADE System Features in Maintein revealed that:

(1) continuing having simple interactions make Maintein easy to use;
(2) the content of prompts needs to be tailored to make not only reflection writing but also conversation starters easier;
(3) further study on frequency, timing and message framing needs to done to be improved the persuasiveness of the reminder and praises of Maintein.

6 FUTURE WORKS

Further studies should be conducted. To improve the experiment, the following should be considered:

(1) The experiment is recommended to have a longer duration. The seven (7) week long period is not enough to really understand the dynamic changes in the attitude and behavior of using Maintein;
(2) The study only had 25 participants, which is a fairly small sample size. The experiment should have more participants and consideration should be taken on their completing the experiment;
(3) All questionnaires need to be revised and piloted to really capture the factors being measured;
(4) The TPB Post Questionnaire should not only be done at the end of the experiment but several times within the duration of the experiment to capture the dynamic nature of changes in attitude and behavior of using Maintein.
(5) The environment and the timing of the use of Maintein by the participants should also be captured and described; and
(6) Succeeding experiments should also consider a third party perspective where the behavior of the participants may not have changed.

Further development of Maintein in terms of its persuasiveness should be done. It should consider the following:

(1) The results of the study as specified in the Section 5;
(2) Include a facility for social support;
(3) Consider having dynamic and personalized content;
(4) Have an option to turn on/off notifications;
(5) Add calm music; and
(6) Randomize questions of the mental well-being scale as some participants feel answering the assessment becomes monotonous and boring.

REFERENCES

[1] Icek Ajzen. 2006. Constructing a Theory of Planned Behavior Questionnaire. , 12 pages.

[2] Zahra Bagheri, Azlina Mohd Kosnin, and Mohammad Ali Besharat. 2016. Improving Emotion Regulation skills through an Emotional Intelligence Training Course. *Khazar Journal of Humanities and Social Sciences* 19 (2016), 36–48. http://dspace.khazar.org/bitstream/20.500.12323/3466/1/Zahra%20Bagheri.pdf

[3] Marc A. Brackett and John D. Mayer. 2003. Convergent, Discriminant, and Incremental Validity of Competing Measures of Emotional Intelligence. *Personality and Social Psychology Bulletin* 29, 9 (2003), 1147–1158. https://doi.org/10.1177/0146167203254596 arXiv:https://doi.org/10.1177/0146167203254596 PMID: 15189610.

[4] Marc A. Brackett, John D. Mayer, and Rebecca M. Warner. 2004. Emotional intelligence and its relation to everyday behaviour. *Personality and Individual Differences* 36 (2004), 1387–1402. https://doi.org/10.1016/S0191-8869(03)00236-8

[5] C.B Cha and MK Nock. 2009. Emotional intelligence is a protective factor for suicidal behavior. *Journal of American Academy of Child and Adolescent Psychiatry* 48, 4 (2009), 422–430. https://doi.org/10.1097/CHI.0b013e3181984f44.

[6] BJ Fogg. 2009. A Behavior Model for Persuasive Design. ePDF. , 7 pages. https://www.mebook.se/images/page_file/38/Fogg%20Behavior%20Model.pdf

[7] Gobinder Singh Gill. 2016. Examining the impact of emotional intelligence and goal setting on basketball performance. *The Sport Journal* 21 (2016). https://thesportjournal.org/article/examining-the-impact-of-emotional-intelligence-and-goal-setting-on-basketball-performance/

[8] Matthew P. H. Kan and Leandre R. Fabrigar. 2017. *Theory of Planned Behavior*. Springer International Publishing, Cham, 1–8. https://doi.org/10.1007/978-3-319-28099-8_1191-1

[9] Walter LaMendola and Judy Krysik. 2008. Design Imperatives to Enhance Evidence-Based Interventions with Persuasive Technology: A Case Scenario in Preventing Child Maltreatment. *Journal of Technology in Human Services* 26, 2-4 (2008), 397–422. https://doi.org/10.1080/15228830802097364 arXiv:https://doi.org/10.1080/15228830802097364

[10] Kathryn Robertson. 2005. Active Listening. *Australian Family Physician* 34, 12 (2005), 1053–1055. https://pdfs.semanticscholar.org/e29b/8c32a6a0788786278e3f4bf6fdd8136d9db7.pdf?_ga=2.24282938.1885720621.1571579415-1895819374.1567161602

[11] Desiree Ruiz-Aranda, Ruth Castillo, Jose Martin Salguero, and Rosario Cabello. 2012. Short- and Midterm Effects of Emotional Intelligence Training on Adolescent Mental Health. *Journal of Adolescent Health* 51, 5 (2012), 462–467. https://doi.org/10.1016/j.jadohealth.2012.02.003

[12] Malouff LE Haggerty DJ Cooper JT Golden CJ Dornheim G Schutte, NS. 1998. Development and Validation of a Measure of Emotional Intelligence. *Personality and Individual Difference* 25 (1998), 167–177.

[13] Oliver Serrat. 2017. *Understanding and Developing Emotional Intelligence*. Springer, 329–339. https://link.springer.com/chapter/10.1007/978-981-10-0983-9_37

[14] Isaac Wiafe and Dorothy Frempong. 2015. Enhancing persuasive features of Behaviour Change Support Systems: The role of U-FADE. *CEUR Workshop Proceedings* 1369 (01 2015), 17–27.

[15] Isaac Wiafe, Keiichi Nakata, and Stephen Gulliver. 2011. Designing Persuasive Third Party Applications for Social Networking Services Based on the 3D-RAB Model. *Communications in Computer and Information Science* 185. https://doi.org/10.1007/978-3-642-22309-9_7

Analyzing the Effects of Subjectivity on Classifying Emotions in Music

Fritz Edron M. Calimag
De La Salle University
Manila, Philippines
fritz_calimag@dlsu.edu.ph

Emir Christopher J. Mendoza
De La Salle University
Manila, Philippines
emir_mendoza@dlsu.edu.ph

Graciel Myka G. Nuncio
De La Salle University
Manila, Philippines
graciel_nuncio@dlsu.edu.ph

Mitchell Bryan Ong
De La Salle University
Manila, Philippines
mitchell_ong@dlsu.edu.ph

Jordan Aiko Deja
De La Salle University
Manila, Philippines
jordan.deja@dlsu.edu.ph

Ryan Austin Fernandez
De La Salle University
Manila, Philippines
ryan.fernandez@dlsu.edu.ph

ABSTRACT

Classifying emotions in music is a complex task due to the subjectivity of emotions. We explore the behavior of classifiers when classifying emotions in music using labels from individual music experts versus labels computed via the consensus of several experts. We performed statistical tests of significance on the precision, recall, and *AUC-ROC score* of classifiers trained across these two sets of labels. Performing the test across all emotions, there was no significant difference in the performance of the classifiers in both setups. Performing the tests with respect to each specific emotion, the classifiers for *calm, bravery, fearful,* and *sadness* performed significantly better when provided labels from the consensus as opposed to labels from a single expert. We provide a research pipeline for testing the effects of subjectivity on the performance of classifiers, as well as recommendations on data collection and representation to improve future work on the topic.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**; **Media arts**; • **Computing methodologies** → *Classification and regression trees.*

KEYWORDS

emotion classification, machine learning, rule mining

ACM Reference Format:
Fritz Edron M. Calimag, Emir Christopher J. Mendoza, Graciel Myka G. Nuncio, Mitchell Bryan Ong, Jordan Aiko Deja, and Ryan Austin Fernandez. 2020. Analyzing the Effects of Subjectivity on Classifying Emotions in Music. In *Workshop On Computation: Theory And Practice, November 21, 2020, Online Conference.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

WCTP 2020, November 21, 2020, Online Conference

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Music is a combination of sounds that harmonize together, and this allows music to embody emotions [16]. Emotions contain underlying meaning discovered from subjective interpretations and understandings of people. Due to this subjective nature, *thematic analysis* is used as a foundational method to capture emotions. However, this method does not have a strict set of rules to follow upon execution. This subjectivity also makes classifying emotions in music difficult [1, 17].

Numerous studies have implemented various techniques in music classification. Current approaches include using regression to predict valence and arousal [18], training multi-modal mood classifiers using both audio and lyric features [7], or detecting moods from acoustic music data using features extracted from the intensity, timbre, and rhythm of the music [13]. These works explored approaches to emotion classification in music, but were focused on predicting *objective* measures of emotion in music such as valence and arousal.

There has been work on specifically classifying categorical, subjective emotion labels using multilayer perceptrons [10]. These emotions were defined as *fearful, cheerful, bravery,* and *love*, using purely low-level musical features as input to the classifier. For our study, ground truth labels were provided by getting the consensus of several experts. Thus, the subjectivity of emotions was not focused on.

Currently, there are limited works on analyzing the effect of subjectivity on the performance of classifiers. There have been works on the subjectivity of identifying specific elements of the music, such as which harmonies are present [11, 12]. Additionally, there have also been studies on how differences in cultural and personal experiences may directly affect the subjective perception of emotions, specifically in music [6, 9]. These works focused on either the qualitative aspect of the emotions in music or on the subjectivity of the specific elements of the music, rather than the overall emotion that can be perceived in the music or subjectivity of emotions in music in classification tasks.

We explore how the subjectivity of emotions in music can affect the performance of classifiers for these emotions. Part of our exploration is identifying if specific emotions in music would be perceived more subjectively than others. Additionally, we explore how a classifier's performance changes when trained on emotions

labeled by different music experts. We present two major experiments: one where the ground truth labels were defined through validation from the individual annotation of the music samples gathered from a music expert and one where the labels were the computed consensus of the experts.

We first provide a preliminary discussion on the high-level and low-level features that we used in the classification tasks, followed by the data collection and preparation method of the study. Afterward, we describe the experiment setup, state our hypotheses, and summarize the results of our experiments. Next, we present the results of the statistical tests, along with the discussion of the results. Finally, we provide our conclusions and recommendations for future work.

2 PRELIMINARIES

Music features are an important aspect to consider when using music for a specific purpose. A single music feature that has been manipulated can have a severe impact on the music piece. Such elements can be classified into either high-level features or low-level features.

High-level features such as *tempo, mode,* and *arousal* provide descriptions for a sophisticated listener to understand music. The *tempo* defines the speed at which a musical piece is played, which is usually measured by beats per minute (BPM). *Mode* is the specific subset of pitches used to write a given musical excerpt [5]. Lastly, *arousal* is defined as an early emotional response towards music [8]. It is difficult to extract such features directly from audio or symbolic representations such as MIDI [3]. For our study, we defined high-level features as categorical values. Tempo was classified into slow or fast tempo. Mode was classified into three categories, which are major, minor, and dissonant. Lastly, arousal was classified into either low or high.

The set of low-level features covered in our study includes the short-time spectrum of segmented audio signals containing different information of the audio sample. Commonly used low-level features are *temporal features, energy features, spectral shape features,* and *perceptual features* [14, 15]. We also define temporal features as low-level features computed from the audio signal frame (e.g. zero-crossing rate and linear prediction coefficients). We define energy features as the energy content of the signal (e.g. root mean square energy of the signal frame, energy of the harmonic component of the power spectrum, and energy of the noisy part of the power spectrum). We define spectral shape features as the shape of the power spectrum of a signal frame: centroid, spread, skewness, kurtosis, slope, roll-off frequency, variation, Mel-frequency cepstral coefficients (MFCCs). Lastly, we define perceptual features as a value computed using a model of the human hearing process (relative specific loudness, sharpness, and spread).

3 METHOD

3.1 Data Preparation

The data preparation stage consists of three main phases which are data collection, expert validation, and data preprocessing.

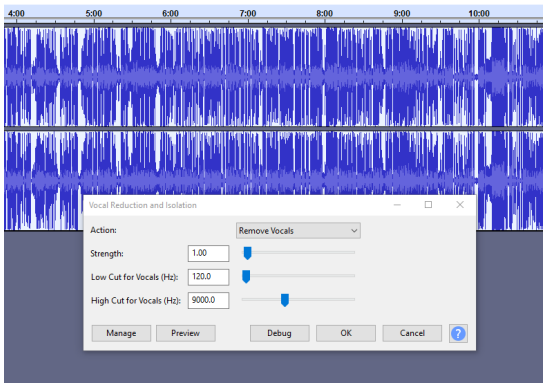


Figure 1: Audacity vocal reduction process. Both audio channels are shown in this screen capture, along with the modal window for removal of vocals.

We collected background music from videos in Chinese children's shows streamed on Youtube, considering YouTube's guidelines that these music samples will be used for educational and non-profit purposes. The keyword used was "Chinese Children's stories" and the majority of these samples were found in the YouTube channel called "Chinese fairy tales". We extracted samples from a total of 57 videos. This type of music was used for scoping purposes and regional relevance. We selected music of one type to minimize the subjectivity from differing cultural experiences as studied by [6, 9]. The longest video reached 30 minutes long while the shortest video was around seven minutes. A third party software called "FFmpeg" was used to convert the videos into audio files which generated 57 files. Each video was extracted to an mp3 file with a sample rate of 4410 Hz and a bitrate of 192 KBPS. Each file had a size of 9 to 20 MB depending on the length of the file. These audio files were separated into samples. The samples were then sorted and temporarily organized into the emotions of *love, cheerful, fearful, sadness, calm*, and *bravery* based on our subjective perception of the emotions present in the audio samples. We were able to collect a total of 421 samples. The resulting dataset from this method of preparation separates this research from other related works.

We used Audacity, an open-source audio editor, to process the music files, separating them into segments and allowing us to separate the narration from the lyrics. Audacity achieves this by first splitting the track into two stereo channels and then inverting the right channel as shown in Figure 1. Inverting the right channel will turn positive waveforms to negative waveforms and vice versa. When the inverted right channel is played together with the left channel, the narration appears to be canceled out. Lastly, the left and right channels were changed to a mono channel. This process, as seen in Figure 2, applies the concept of vocal reduction and isolation, which tries to remove the center-panned audio from stereo tracks. We chose this approach since vocals are mostly recorded in this manner and it is one of the simplest methods in removing vocals. However, for some files, the vocals may not be recorded in a manner suitable for this process.

After removing the vocals, certain segments of the samples which represent a certain emotion are selected for clipping. The resulting

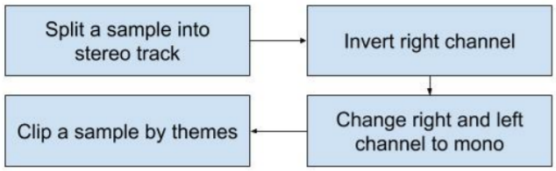


Figure 2: Summary of vocal reduction process in Audacity [10]

Table 1: Extracted Low-Level Features

Area Method of Moments of MFCCs	Beat Sum
Compactness	Fraction Of Low Energy Windows
LPC	Method of Moments
Root Mean Square	Spectral Centroid
Spectral Flux	Spectral Rolloff Point
Spectral Variability	Strength Of Strongest Beat
Strongest Beat	Zero Crossings

clipped segments are exported as a .wav file with a sample rate of 4410 Hz and a bit rate of 1411 kbps. After the audio was clipped and exported, we produced 75 audio samples for each emotion, excluding *sadness* which only had 46, totaling up to 421 samples. However, 2 samples were labeled as "broken samples" during the pre-processing and were removed from the final list of samples. This resulted in a final count of 419 samples which were made ready for expert validation.

Music is affected by different features which can greatly affect the emotion a sample can convey. This informs the need for experts to label our samples as they can properly differentiate the features used in each sample. We chose three music experts that had at least five years in the field of music and had an educational background in music. These experts were asked to identify the tempo, arousal, mode, and emotion of the samples as defined in Section 2. The samples were randomly arranged and were given non-leading file names such as "001.mp3", allowing the experts to label them without prior bias.

The audio samples were placed into JAudio after it was segregated. JAudio is the feature extraction tool used to get the low-level features. Each output was an XML file that has an attribute name and value about each sample, which was then converted into a CSV file. Table 1 shows some of the extracted features.

We performed min-max scaling on our dataset after we finished extracting the features using JAudio. We split our dataset into a 95% training set and 5% validation set. The validation set was randomly selected in a way such that each of the six emotions was almost uniformly represented. All of the samples in the validation set were properly labeled.

Some of the samples were left unlabeled due to either limited availability of the experts or complete disagreement of the three experts (these limitations are further discussed in Section 4.3). We isolated all of these samples to the 95% training set. To address the missing labels, we used a semi-supervised learning approach. We first trained classifiers on the properly labeled samples in the training set using the various machine learning algorithms described in section 3.2. The best performing classifier based on f1-score, which was the Decision Tree model, was used to label the unlabeled samples in the training set. Combining these two sets led to the final training set.

Feature selection was done by first removing low-variance features followed by applying a χ^2 feature ranking and selecting the k-best features. Afterward, the dataset was found to be imbalanced in terms of the samples for each emotion. To address the imbalance, we used random oversampling and Synthetic Minority Oversampling Technique (SMOTE) [4] to generate two additional versions of the dataset.

3.2 Experiment Setup

We performed experiments to answer the following research questions:

- How do classifiers respond to the subjectivity of emotions in music?
- For each emotion, how do the classifiers respond to the subjectivity of the labels?
- How did the classifiers handle the subjectivity of emotions in music across several experts?

For each of these, the following hypotheses were tested respectively (for brevity, the labels extracted via consensus of the experts are referred to as *consensus labels* for the rest of the document):

- H_1 : There is a significant difference in a classifier's performance when trained on a single expert's labels as opposed to the consensus labels.
- H_2 : For each emotion, there is a significant difference in a classifier's performance when trained on a single expert's labels as opposed to the consensus labels.
- H_3 : There is a significant difference in a classifier's performance across the labels of each of the three experts.

For comparing the results of various sets of experiments, we performed statistical tests of significance with our significance level set to $p < 0.05$.

Two main experiment setups were conducted: one where the classifiers were trained on the labels from a single expert only and one where the classifiers were trained on the consensus labels. For each set of labels, several classifiers were trained: one for each combination of algorithm, dataset (imbalanced, random oversampling, or SMOTE), and emotion. This means each classifier would be a binary classifier that identifies a single emotion.

To build the classifiers, the following machine learning algorithms were selected:

- AdaBoost
- Decision Tree
- Gradient Boost
- K-Nearest Neighbors
- Logistic Regression
- Multilayer Perceptron
- Naïve Bayes Classifier

Table 2: Summary of mean of our performance measures of the two experiments. Experts are anonymized as E1, E2, and E3. We report the mean precision, recall, and AUC-ROC scores of each group of trained classifiers. Maximum values are displayed in bold. Minimum values are underlined

Emotion	Expert	Mean Precision	Mean Recall	Mean AUC-ROC Score
Calm	E1	0.60	<u>0.61</u>	0.60
	E2	0.64	0.73	0.52
	E3	<u>0.61</u>	0.74	0.48
	Consensus	0.67	0.70	0.50
Cheerful	E1	0.81	0.86	0.62
	E2	0.72	0.79	0.47
	E3	0.80	0.87	0.62
	Consensus	0.73	0.85	0.50
Bravery	E1	0.77	0.83	0.52
	E2	0.72	0.79	<u>0.46</u>
	E3	0.73	0.83	0.50
	Consensus	0.73	0.77	0.55
Fearful	E1	0.69	0.80	0.54
	E2	0.63	0.69	0.48
	E3	<u>0.61</u>	0.71	<u>0.46</u>
	Consensus	0.72	0.77	0.56
Love	E1	0.85	**0.90**	0.49
	E2	0.73	0.77	0.49
	E3	0.67	0.75	0.54
	Consensus	0.73	0.83	0.50
Sadness	E1	**0.86**	0.87	<u>0.46</u>
	E2	0.73	0.83	0.48
	E3	0.83	0.81	0.44
	Consensus	0.85	0.86	**0.71**

- Random Forest
- Regularized Linear Model
- Rule-Based Classifier as described in [2]
- Support Vector Machine

We used a random hyperparameter search with k-fold cross-validation to train the classifiers with $k = 5$. The final classifier was then run on the 5% validation set. To measure the performance of the classifiers, we recorded the *precision, recall*, and *AUC-ROC scores*.

4 RESULTS AND ANALYSIS

4.1 Summary of Results

Table 2 shows a summary of the results of our experiment. The three experts have been anonymized as E1, E2, and E3. In terms of *precision*, the classifiers of *sadness* for E1 performed the best on average. In terms of *recall*, the classifiers of love for E1 performed the best on average. In terms of *AUC-ROC score*, the classifiers of *sadness* for the consensus labels performed the best. Conversely, in terms of *precision*, the classifiers of *calm* for E3 performed the worst on average. In terms of *recall*, the classifiers of *calm* for E1 performed the worst on average. In terms of *AUC-ROC score*, the classifiers for *bravery* for E2, *fearful* for E3, and *sadness* for E1 performed the worst on average.

4.2 Discussion

Returning to the research questions and hypotheses posed earlier in this paper, we conducted the statistical tests of significance and arrived at the following conclusions:

A two-tailed Student's t-test was done to test whether there is a significant difference in a classifier's performance when trained on a single expert's labels as opposed to the consensus labels, specifically whether there was a significant difference between the means of the performance measures of the classifiers trained on single expert labels versus those trained on the consensus labels (H_1). For *precision*, we concluded that there was no significant difference between the two means ($t = 0.65, p = 0.51 \geq 0.05$). For *recall*, we also concluded no significant difference between the two means ($t = -1.25, p = 0.21 \geq 0.05$). Finally, for the *AUC-ROC scores*, we also concluded no significant difference between the two means ($t = 1.08, p = 0.27 \geq 0.05$). These test results indicate that, when considering all emotions, the classifier's performance was not greatly affected by whether the ground truth labels came from the opinion of a single expert or a consensus. This implies that, in general, subjectivity does not have a strong effect on the performance of a classifier.

Similarly to H_1, to test whether, for each emotion, there is a significant difference in a classifier's performance when trained on a single expert's labels as opposed to the consensus labels (H_2), a two-tailed Student's t-test was performed. For this test, however, the classifiers were segregated into six groups, one for each emotion.

Proceeding to the tests of H_2, Table 3 summarizes the t-statistics and p-values of the tests. We concluded that there were three emotions wherein the classifier performed better in terms of *precision* using the consensus labels compared to the individual expert labels: *calm*, *fearful*, and *sadness*. For this measure and these emotions, the results support H_2. Similarly, the classifier performed better in terms of *AUC-ROC score* using the consensus labels for *bravery, fearful,* and *sadness*, likewise supporting H_2. The classifiers trained on the rest of the emotions had no significant differences in terms of *precision* and *AUC-ROC score*, notably. The classifiers had no significant differences in performance in terms of *recall* in all six emotions.

These test results indicate that some combinations of measures and emotions are more sensitive to subjectivity than others. Of note are the emotions of *fearful* and *sadness*, which perform better on consensus labels for two out of the three performance measures. This would match with prior assumptions since the emotions of fear and sadness could be said to be highly subjective to personal culture and experiences, as discussed in [6, 9]. Since the experts come from similar backgrounds, it may be possible that the consensus labels stem from a commonly accepted notion of fear or sadness across all the experts. Concerning the related works, our dataset uses a different set of labels, therefore comparison was not possible.

Table 3: Summary of p-values for statistical tests for H2. We performed a two-tailed Student's t-test between the set of all classifiers trained on a single expert's labels and the set of all classifiers trained on consensus labels. Significance level of the test was set to $p < 0.05$. P-values that indicated a significant difference are displayed in bold.

Emotion	Measure	t-statistic	p-value
Calm	Precision	-2.08	**0.02**
	Recall	-0.33	0.36
	AUC-ROC Score	1.19	0.11
Cheerful	Precision	1.61	0.05
	Recall	-0.29	0.35
	AUC-ROC Score	1.36	0.09
Bravery	Precision	0.30	0.38
	Recall	1.67	0.05
	AUC-ROC Score	-2.63	**0.00**
Fearful	Precision	-3.66	**0.00**
	Recall	-0.95	0.17
	AUC-ROC Score	-2.68	**0.00**
Love	Precision	0.62	0.26
	Recall	-0.81	0.21
	AUC-ROC Score	-0.81	0.46
Sadness	Precision	-1.70	**0.04**
	Recall	-0.88	0.19
	AUC-ROC Score	-7.84	**0.00**

Finally, to test whether there is a significant difference in a classifier's performance across the labels of each of the three experts (H_3), we performed a one-way ANOVA test across three populations, which were the groups of classifiers trained on the labels of each of our three experts. For *precision*, we concluded that there is no significant difference in the mean performance across all three sets of classifiers ($F = 2.84, p = 0.05 \geq 0.05$). For *recall*, we also concluded that there is no significant difference in the mean performance across all three sets of classifiers ($F = 1.58, p = 0.20 \geq 0.05$). However, for *AUC-ROC scores*, we concluded that there is a significant difference in the mean performance across all three sets of classifiers ($F = 15.39, p = 0.00 < 0.05$). This indicates that, in general, the classifiers were able to learn consistently well, relative to one another, regardless of the source of the label, but some of the classifiers may be responding to experts who are more consistent in identifying emotions in music as compared to the other experts from a computational standpoint. However, the *AUC-ROC scores* generally had values of 0.5 or lower, as seen in Table 2. The significant difference might be caused by a few instances of the data having *AUC-ROC scores* of 0.7 - 1.0. These results would indicate though that some trained classifiers still perform better if the annotator of the emotions is more consistent with how they label their emotions.

4.3 Limitations

Due to the circumstances brought about by the COVID-19 pandemic, one of our experts was not able to finish labeling the entire dataset. As mentioned in Section 3.1, we proceeded with a semi-supervised learning approach. We also ended up with an imbalanced dataset, using oversampling techniques to arrive at a balanced dataset [4]. Additionally, only three experts were contacted for the study, which means all experiments only deal with the labels and consensus of three experts.

Algorithms such as Naïve Bayes perform better using categorical data, which our dataset has a limited representation of. Classifiers such as these mainly use high-level music features in the context of our study. Having only 12 possible combinations for these, the representation was insufficient for the classifiers to properly classify the samples. This is a possible indicator that the distribution of our data only properly utilizes our low-level music features. Even with Gaussian Naïve Bayes classifier, which is meant to deal with continuous data, the algorithm garners lower results than the rest of the algorithms.

5 CONCLUSIONS AND FUTURE WORK

We explored the effects of the subjectivity of human emotion on classifying emotions in music. We presented a research pipeline that tested whether the labeling method of a dataset where emotions are the ground truth would affect the performance of the classifiers. After performing the statistical tests with the *precision, recall,* and the *AUC-ROC scores*, we found that there was no significant difference between classifiers that were trained using an individual expert's labels and those trained using the consensus expert labels, implying that subjectivity, in general, does not have a significant effect on classification tasks. However, for some emotions, such as *fearful* and *sadness*, the classifiers trained on consensus labels performed significantly better than those trained on single expert labels. This may indicate that these emotions are strongly dictated by subjective cultural experiences as opposed to personal experiences. Finally, there was a significant difference in the performance of the classifiers across the three experts, possibly indicating that some of the experts were more consistent with their labeling of

the emotions and that the classifiers are sensitive to this internal consistency within sets of labels. Overall, our results indicate that for some of the emotions, the classifiers to respond to the subjective nature of some of the ground truth labels, but in general, the algorithms we chose are robust enough that subjectivity does not have a significant effect on its performance.

Future work on this topic can expand on our work with the following recommendations: first, the gathering of the samples can be improved upon; some of the final music samples ended up having white noise that would affect the emotion of the music sample. There were also music samples that ended up being labeled as a broken sample as its music content was omitted in the process of cleaning the data. We suggest using a different approach in removing the narration for cleaning the data to avoid any of these anomalies.

Secondly, would be the addition of a middle-range in tempo during expert labeling, for we were advised by the musical experts during data validation that tempo is not only restricted between slow and fast. The music experts suggested giving a range of beats per minute on what is fast and slow.

Finally, our conclusions were made with the experiment only considering a total of three musical experts. Having more musical experts could provide a clearer consensus on what emotion a sample classifies into. This could improve the balancing of the dataset and could also improve the results and provide more meaningful insights on the effect of subjectivity on the performance of classifiers of emotions in music.

REFERENCES

[1] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic analysis. *Handbook of Research Methods in Health Social Sciences* (2019), 843–860.

[2] Anna L Buczak, Phillip T Koshute, Steven M Babin, Brian H Feighner, and Sheryl H Lewis. 2012. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC medical informatics and decision making* 12, 1 (2012), 124.

[3] Michael A Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. 2008. Content-based music information retrieval: Current directions and future challenges. *Proc. IEEE* 96, 4 (2008), 668–696.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[5] Simone Dalla Bella, Isabelle Peretz, Luc Rousseau, and Nathalie Gosselin. 2001. A developmental study of the affective value of tempo and mode in music. *Cognition* 80, 3 (2001), B1–B10.

[6] Nicola Dibben. 2006. Subjectivity and the Construction of Emotion in the Music of Björk. *Music Analysis* 25, 1-2 (2006), 171–197. https://doi.org/10.1111/j.1468-2249.2006.00237.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2249.2006.00237.x

[7] Xiao Hu and J Stephen Downie. 2010. When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis.. In *ISMIR*. 619–624.

[8] Ericka Kelley, Gabrielle Andrick, Fayelin Benzenbower, and Marlene Devia. 2014. Physiological arousal response to differing musical genres. *Modern Psychological Studies* 20, 1 (2014), 4.

[9] Annekatrin Kessler and Klaus Puhl. 2004. Subjectivity, emotion, and meaning in music perception. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM04) Graz/Austria*. Citeseer, 15–18.

[10] YuSin Kim. 2019. *Building A Model for Music Classification in Children's Stories Using Neural Network*. Master's thesis. De La Salle University Manila.

[11] Hendrik Vincent Koops, W. Bas de Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller, and Anja Volk. 2019. Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research* 48, 3 (2019), 232–252. https://doi.org/10.1080/09298215.2019.1613436 arXiv:https://doi.org/10.1080/09298215.2019.1613436

[12] Hendrik Vincent Koops, W Bas de Haas, John Ashley Burgoyne, Jeroen Bransen, and Anja Volk. 2017. Harmonic subjectivity in popular music.

[13] Lie Lu, Dan Liu, and Hong-Jiang Zhang. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing* 14, 1 (2006), 5–18.

[14] May Thu Myint. 2019. A Classification and Comparison of Feature Extraction For Myanmar Ethnic Songs. (2019).

[15] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek. 2006. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine* 23, 2 (2006), 133–141.

[16] Mojtaba Vaismoradi, Jacqueline Jones, Hannele Turunen, and Sherrill Snelgrove. 2016. Theme development in qualitative content analysis and thematic analysis. (2016).

[17] Yi-Hsuan Yang and Homer H Chen. 2011. *Music emotion recognition*. CRC Press.

[18] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. 2008. A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing* 16, 2 (2008), 448–457.

Epidemiological Network Analysis for COVID-19 Contact Tracing

Adrian Jose Sabado, Alfred John Tacorda, Elias Andre Chico, Kenneth Antonio, Jean Nathan Cabreza and Geoffrey Solano
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
adrian310sabado@gmail.com, agtacorda@up.edu.ph, danerprog@gmail.com, jgantonio1@up.edu.ph, jdcabreza@up.edu.ph, gasolano@up.edu.ph

Monico Galicia, Dominique Torralba, Elijah Puaben and Christian Edmund Chua
SGV & Co.
Manila, Philippines

***Abstract*—Contact tracing is one of the critical tools in managing the transmission of COVID-19. According to the WHO guidelines, it is applicable for areas with zero or sporadic cases, all the way to areas with clusters of cases. However, information problems are encountered as contact tracing is done at the field level. This study was developed to support the local government's efforts towards more efficient contact tracing as well as the reduction of information gaps by augmenting the contacts' responses with their mobile phone's GPS history to create a contact-location network. Analysis is then performed on the network by extracting centrality measures (degree, weighted degree betweenness, closeness, eigenvector) to identify high-risk targets for monitoring as well as location hotspots. The output may also be used to validate results of field worker investigations.**

***Index Terms*—contact tracing, COVID-19, network, centrality**

I. Introduction

COVID-19 is a disease related to SARS-CoV-2, a virus which initially broke out in Wuhan, China was officially declared as a global pandemic by the WHO on January 30, 2020 [1]. COVID-19 is highly transmittable [2] and infection mainly happens due to close contact with an infected person [3]. Contact tracing is an effective way to stem COVID-19 transmission [4, 5] since it enables early identification and monitoring of people who have had close contact with an infected person. It has effectively been applied in several locations internationally, both on a local scale [7] and a country-wide scale [6].

However, there exists a few problems with contact tracing at the field level. Patients often forget (or incorrectly recall) their traveling history and previous contacts. They also usually only provide very general or high-level information. These problems may make manual contact tracing difficult and inaccurate. Phone GPS systems have the ability to record both history and location without the need for their users to manually log data and so make for a good record of a persons' travel history and possible contact points with other people.

Thus the study presented in this paper was embarked upon by SGV & Co., along with the Department of Science and Technology and the Mathematical and Computing Sciences Unit of UP Manila. The aim is to further examine the said data through network analysis to provide essential information on other vulnerable contacts as well as locations that need to be monitored by local government units (LGUs). These can also provide information for epidemiologists to study patterns of transmission to identify sources of contagion.

II. Related Work

A. Contact Tracing

Contact tracing is the process of identifying, assessing, and managing individuals who have been exposed to a disease with the aim of preventing onward transmission. When systematically applied, contact tracing will break the flow of transmission of an infectious disease and is thus an essential public health tool for controlling infectious disease outbreaks [8, 9, 19]. Contact tracing preparedness is necessary across different epidemiological scenarios whether in zero cases, sporadic cases, clusters or even in instances of community transmission. Even when countries have passed the peak of transmission and case numbers are dwindling, rapid identification of cases and contact tracing are critical [19].

It has previously been used on SARS data by Chen et al. [10], another coronavirus that is very closely related to SARS-CoV-2. Chen et al. also used networks, but they had also included geographical data in addition to the social network. This is very similar to the approach used in this study in that the GPS automatically encodes geographical locations in the data (and by extension, the network).

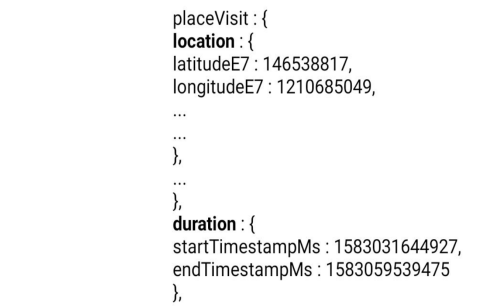


Fig. 1. Example of Google location history data downloaded using AGAP

In addition to SARS data, it has also been shown that contact tracing may have been effective on managing the spread of

Fig. 2. Screenshots of the AGAP mobile app

smallpox [11] and it has also been used in analyzing the spread of sexually transmitted diseases [12].

B. Contact Tracing for COVID-19

Contact tracing for COVID-19 requires identifying persons who may have been exposed to COVID-19 and following them up daily for 14 days from the last point of exposure. A document [19] containing guidelines on how to establish contact tracing capacity for the control of COVID-19 has been released by the World Health Organization(WHO).

There exists several (both proposed and deployed) COVID-19 tracing systems [13], but there also exists several problems with privacy and data [14]. Some of applications may ask for more permissions than they require and they may be outsourcing data to unknown third parties [15]. This, therefore, makes it hard for local governments to adopt this kind of technology despite their availability.

C. Network Analysis for Contact Tracing

Network analysis is a method that has been used in studies such as contact tracing for tuberculosis in Vietnam [16] and for SARS data from Taiwan [10]. A person at the center of a network is more susceptible to the risks of social connection than those at the periphery of a network. People are thus affected by their location in a social network.

Another study [17] has also performed contact tracing via network analysis and centrality measures for COVID-19. They used both degree and betweenness centrality to analyze a directed graph created from COVID-19 data collected in India. We use other additional centrality measures to provide a more complete view of the relationships between nodes. Furthermore, the network model used in this study we do not use directed graphs a well since contact time is a two-way relationship.

III. Methodology

A. Data Collection

The data used for contact tracing is the Google location history data, which are GPS coordinates captured in mobile phones of people who have consented to contact tracing from the local government. This is facilitated by the **UP Manila AGAP** which is a utility app designed to help in the government's efforts in preventing the further spread of Covid-19. AGAP makes use of the Google Takeout service to obtain one's location history data. An example of example of location history data downloaded using AGAP is seen in Figure 1.

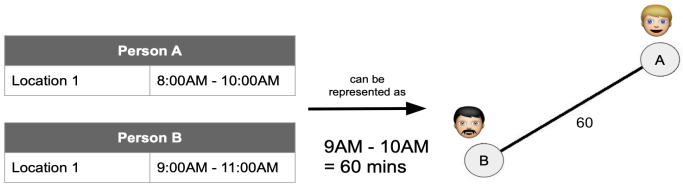


Fig. 3. Example graph representation of two people who have come in contact with each other

When a contact is interviewed by a contact tracing officer at the field level, the following information problems are usually encountered: 1.) patients forget or incorrectly recall details of their previous contacts and travel history for more than a few days, and 2.) patients provide only very general or high-level information. These problems will be avoided with this contact tracing approach that makes use of the AGAP mobile application, as the contact will be asked for consent if he/she is willing to share his/her GPS history. The contact is then instructed to download AGAP. All logins and authentications are done using Google's services. Only the location history for the past 14 days is taken from Google Takeout. The data

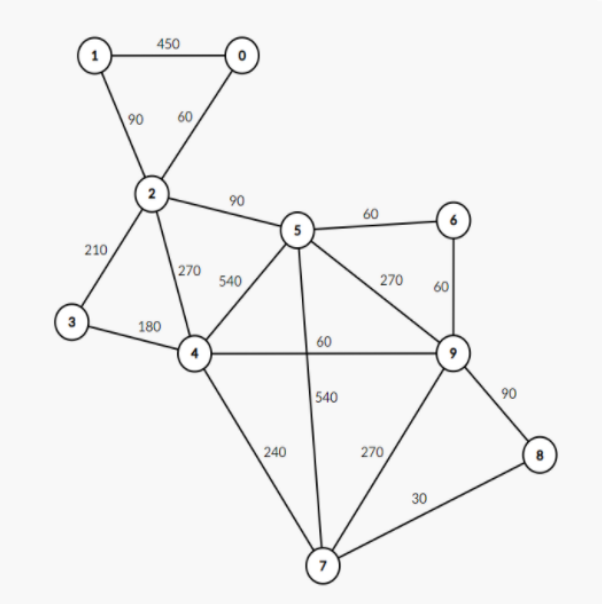


Fig. 4. A network of 10 different people with corresponding total contact duration

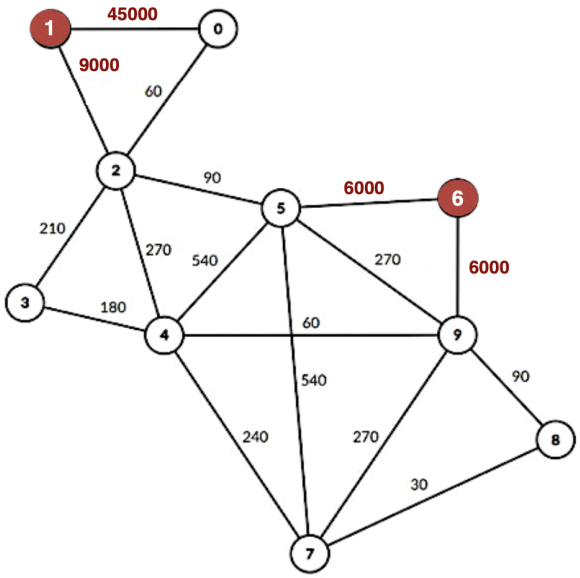


Fig. 5. Edge weights are modified when a case is confirmed.

is then shared with the local government unit. Screenshots of the AGAP mobile app are seen in Figure 2.

Aside from the network that will be build for contact tracing, the data obtained may be used to identify Covid-19 hotspots. Frequency counts were made to determine locations that were most visited as well as high contact places.

B. Building the Network

Each person is represented by a node in the network. An edge exists between two nodes if the corresponding individuals visited the same place at the same time. Edges are also given weights, which represent the length of time that two people are in the same place at the same time (i.e. contact time between the two people), as shown in Figure 3.

The resulting graph is an undirected graph since contact is a two-way behavior (i.e. contact is mutual between two people). In case the same couple of people are in another venue for a given duration, say 1.5 hours, then 90 is added to the weight of the edge incident to them. A network of 10 different people with corresponding total contact duration is shown in Figure 4.

We also note that the network is dynamic. New nodes and edges are added as individuals are interviewed and their corresponding GPS histories are downloaded. Edges may also be deleted if contact has occurred much in the past, e.g. more than 30 days past.

Edge weights are also dynamic. Whenever a person becomes a confirmed COVID-19 case, the weights of the edges incident to the corresponding node is multiplied by a constant ρ, which is decided upon by the network analysts. This will therefore have an effect on the centrality measures of neighboring nodes. In Figure 5, if nodes 1 and 6 have become confirmed cases, the weight of all the edges incident to either node 1 and node 6 are multiplied with a large value $\rho = 100$. This ensures that people who become confirmed cases remain with higher centrality measures while the people who they have come into contact with will also experience an increase in the values of their own centrality measures, making it easier to determine who will be needing to be tested and contact traced further.

C. Centrality Measures

Centrality is the measure of importance of nodes in the network. Now, since the concept of importance is ambiguous, there are a variety of centrality measures used.

The following centrality measures will be used to examine and interpret the network of people generated by some given data: degree centrality, weighted degree centrality, betweenness, closeness, and eigenvector centrality. Each are further discussed in the subsections that follow.

1) Degree Centrality: Degree centrality is defined as the number of edges that are incident (that is, connected) to a particular node. A higher degree centrality means that the node is more central, that is, it has some significant amount of connections to other nodes.

2) Weighted Degree Centrality: Weighted degree centrality is defined as the sum of the weights of all the edges incident to a particular node. Note that a high weighted degree centrality may not necessarily mean that a node is connected to a large amount of other nodes. It may mean that one of the edge weights connected to a certain node has a significantly large value that inflates the weighted degree centrality of a node.

3) Betweenness: The betweenness centrality of a node refers to the number of shortest paths from every vertex to

Node	Degree Centrality	Weighted Degree Centrality
0	2	459
1	2	486
2	5	648
3	2	207
4	5	345
5	5	222
6	2	12
7	4	108
8	2	12
9	5	75

TABLE I
DEGREE AND WEIGHTED DEGREE CENTRALITY

every other vertex that pass through that particular node. It is given by the formula:

$$C_b(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}} \tag{1}$$

where g_{jk} is the number of shortest paths connecting two nodes j and k and $g_{jk}(i)$ is the number of shortest paths between j and k that the node i is on.

4) Closeness: The closeness centrality of a node refers to the average shortest path between the node and the rest of the nodes in the network. It is given by the formula:

$$C_c(i) = \frac{1}{\sum_{j=1}^{N} d(i,j)} \tag{2}$$

where $d(i,j)$ is the length of the shortest path between a node i and j and N is the total number of nodes in the network.

5) Eigenvector Centrality: The eigenvector centrality of a node refers to how 'influential' (i.e. how large its other measures of centrality are compared to the other nodes in the network) a node is in the network. Every node is assigned an eigenvector centrality that reflects how connected they are to highly influential nodes. Connections to highly influential nodes have a greater effect on the eigenvector centrality (i.e. they significantly increase it) of a given node. The eigenvector centrality is given by:

$$C_e(v) = \frac{1}{\lambda} \sum_{t \in M(v)} C_e(t) = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} C(t) \tag{3}$$

where λ is some constant, $M(v)$ are the neighbors of a node v and $a_{v,t}$ is a value in the adjacency matrix A of the graph G that returns the value of A at $[v,t]$.

IV. RESULTS AND DISCUSSION

A. Sample Experiments

Experimental results were done in this study. For ethical considerations, these are the data results that are presented in this paper. Various updates were performed in the network shown Figure 3 and the resulting network configuration is shown in Figure 6, from which centrality measures were obtained.

It can be seen in the tables shown that different nodes that are presented as critical with respect to different centrality

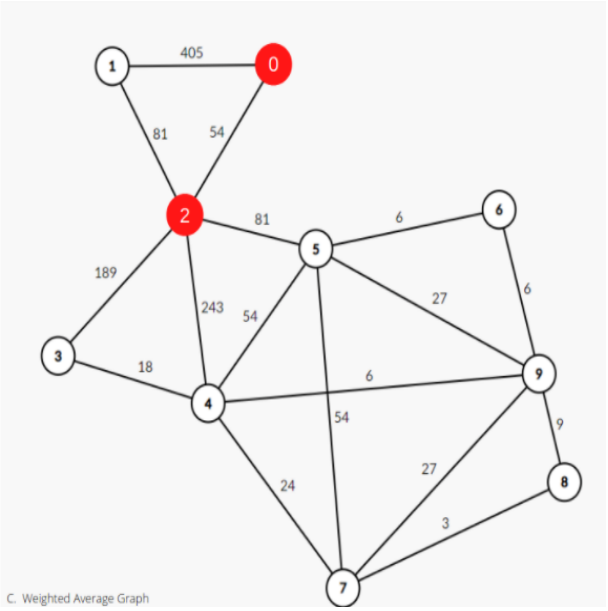


Fig. 6. Resulting network after various updates were performed on the network in Figure 2.

Node	Betweenness	Closeness	Eigencentrality
0	0	0.0072	0.5987
1	0	0.0061	0.6115
2	0.41667	0.0110	0.4063
3	0	0.0136	0.1801
4	0.2222	0.0173	0.2407
5	0.5	0.0182	0.1058
6	0.5556	0.0186	0.0016
7	0	0.0160	0.0262
8	0.2222	0.0168	0.0004
9	0.6667	0.0186	0.0115

TABLE II
BETWEENNESS, CLOSENESS AND EIGENVECTOR CENTRALITY

measures. In terms of degree centrality, nodes 2, 4, 5 and 9 are most central. Nodes 0, 1 and 2 are the most central in terms of weighted degree centrality. Nodes 2, 5, 6 and 9 are the most central in terms of betweenness. On the other hand, for closeness centrality, nodes 5, 6 and 9 are the most central. Finally, in terms of eigenvector centrality, nodes 0, 1 and 2 are the most central.

B. Interpretation of Results

1) Degree Centrality: By the definition of the degree centrality, we can interpret it as the number of people who made contact with a particular node. Nodes with a relatively large degree centrality may be considered as potential super-spreaders. This can serve as aid in gauging the number of people possibly affected by a single person.

2) Weighted Degree Centrality: We interpret the weighted degree centrality as the total time a person made contact with other people. Note that this is not total contact time

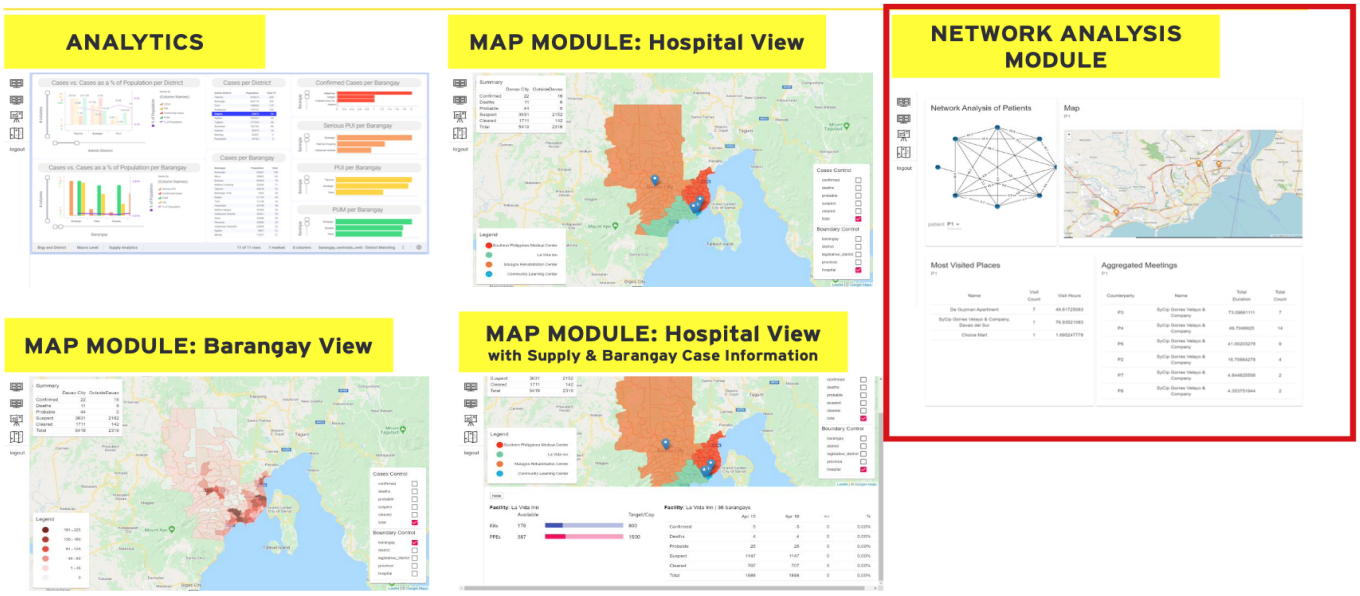


Fig. 7. The dashboard presented to the LGU. The red box shows the network analysis module of the project.

with a particular person, as that is already represented by the edge weight. A large weighted degree centrality, along with a large degree centrality, may further confirm the presence of a potential super-spreader. A node may still exhibit a large weighted degree centrality without a large degree centrality in the case of a small amount of individuals spending significant amounts of time in contact. Therefore, the weighted degree centrality may also be used to gauge the likelihood of a node getting infected, should they make contact with a confirmed infection for a significant amount of time.

3) Betweenness: Betweenness describes how much a specific node acts as a sort of 'bridge' between other nodes. It also means that the corresponding person has the tendency to function as a link or bridge among clusters of cases. This person has contact with various different independent groups and has the potential to spread the virus to these groups.

4) Closeness: We interpret closeness as the degree to which a node is exposed to the outside world. The corresponding person is close or highly accessible or exposed to the whole population which makes it vulnerable.

5) Eigenvector Centrality: Since eigenvector centrality takes into consideration other centrality values, one interpretation for this is that the corresponding person who has high eigenvalue centrality has been highly exposed to other individuals who are also highly exposed.

C. The Project Dashboard

The system was initially deployed to an LGU which we are not at liberty to mention in this study. Figure 7 shows the dashboard that was presented to the LGU. The red box shows the Network Analysis Module of the project.

The Network Analysis Module of the dashboard has the following features :

- The **Map** gives a graphical representation of the participant's location history: Most Visited Places and High Contact Places.
- The **Frequent Locations (FL) Report** shows the individual's visited places ranked according to frequency of visits and duration of visit.
- The **Probable Contact (PC) Report**, obtained from network analysis results, shows the persons in the network with contact with the selected participant, ranked in order of decreasing exposure, together with corresponding location of most frequent contact, duration of contact, and frequency of contact. This can be used to identify next-gen individuals for investigation.

D. Augmenting the Current System

The approaches presented in this study can potentially augment the current contact tracing efforts of LGUs and produce better results.

Here are some ways this study may improve the current system :

- During the conduct of the interview with an identified contact, the Google location history can provide additional detailed data that is accurate. It can also help the contact remember more information.
- When the contact tracing report is generated, the Frequent Location and Probable Contact Reports generated from Google location data as well as network analysis results can provide additional data that contact tracers will find helpful. FL and PC will identify frequented locations not disclosed during the interview and possible contact with existing confirmed cases.

V. Conclusion

This study presents an approach that makes use of Google location history as well as network analysis to provide a more accurate list of individuals that need to be contacted as well as critical locations that need to be monitored. It is hoped that this approach will help contact tracers in LGUs to have more accurate and detailed information as they perform their difficult task. The results can also potentially provide information for epidemiologists to study patterns of transmission to identify sources of contagion.

Acknowledgement

The authors would like to thank SGV & Co. and the Department of Science and Technology (DOST) for the opportunity to work on this project. Thanks also to the Department of Physical Sciences and Mathematics, University of the Philippines Manila for the support and to the UP Manila Computer Science students and alumni for your willingness to help. Special thanks also to the following : Princess Danielle Florendo, Kharl Gaebriel Agir,Marc Jermaine Pontiveros, Kevin Anthony Sison, Andrei Mikail Macatangay and Reuben Joseph Cabrera for being part of this project.

References

[1] Velavan, T. P., and Meyer, C. G. (2020). The COVID-19 epidemic. Tropical medicine & international health, 25(3), 278.

[2] Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., and Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. Journal of Advanced Research.

[3] Coronavirus disease (COVID-19): How is it transmitted? (n.d.). Retrieved from https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted.

[4] Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic?. The Lancet, 395(10228), 931-934.

[5] Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., ... and Eggo, R. M. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health.

[6] Park, Y. J., Choe, Y. J., Park, O., Park, S. Y., Kim, Y. M., Kim, J., ... and COVID-19 National Emergency Response Center, Epidemiology and Case Management Team. (2020). Contact tracing during coronavirus disease outbreak, South Korea, 2020. Emerging infectious diseases, 26(10), 2465-2468.

[7] Lash, R. R., Donovan, C. V., Fleischauer, A. T., Moore, Z. S., Harris, G., Hayes, S., ... and Samoff, E. (2020). COVID-19 contact tracing in two counties—North Carolina, June–July 2020. Morbidity and Mortality Weekly Report, 69(38), 1360.

[8] Klinkenberg, D., Fraser, C., and Heesterbeek, H. (2006). The effectiveness of contact tracing in emerging epidemics. PloS one, 1(1), e12.

[9] Eames, K. T., and Keeling, M. J. (2003). Contact tracing and disease control. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1533), 2565-2571.

[10] Chen, Y. D., Tseng, C., King, C. C., Wu, T. S. J., and Chen, H. (2007, May). Incorporating geographical contacts into social network analysis for contact tracing in epidemiology: a study on Taiwan SARS data. In NSF Workshop on Intelligence and Security Informatics (pp. 23-36). Springer, Berlin, Heidelberg.

[11] Eichner, M. (2003). Case isolation and contact tracing can prevent the spread of smallpox. American journal of epidemiology, 158(2), 118-128.

[12] Garnett, G. P., and Anderson, R. M. (1993). Contact tracing and the estimation of sexual mixing patterns: the epidemiology of gonococcal infections. Sexually transmitted diseases, 20(4), 181-191.

[13] Ahmed, N., Michelin, R. A., Xue, W., Ruj, S., Malaney, R., Kanhere, S. S., ... and Jha, S. K. (2020). A survey of covid-19 contact tracing apps. IEEE Access, 8, 134577-134601.

[14] Wen, H., Zhao, Q., Lin, Z., Xuan, D., and Shroff, N. (2020, October). A study of the privacy of covid-19 contact tracing apps. In International Conference on Security and Privacy in Communication Systems (pp. 297-317). Springer, Cham.

[15] Azad, M. A., Arshad, J., Akmal, S. M. A., Riaz, F., Abdullah, S., Imran, M., and Ahmad, F. (2020). A First Look at Privacy Analysis of COVID-19 Contact Tracing Mobile Applications. IEEE Internet of Things Journal.

[16] Hoang, T. T. T., Nguyen, V. N., Dinh, N. S., Thwaites, G., Nguyen, T. A., van Doorn, H. R., ... and Wertheim, H. F. (2019). Active contact tracing beyond the household in multidrug resistant tuberculosis in Vietnam: a cohort study. BMC public health, 19(1), 241.

[17] Nagarajan, K., Muniyandi, M., Palani, B., and Sellappan, S. (2020). Social network analysis methods for exploring SARS-CoV-2 contact tracing data. BMC medical research methodology, 20(1), 1-10.

[18] Christakis NA, Fowler JH. Social Network Visualization in Epidemiology. Nor Epidemiol. 2009;19(1):5-16.

[19] World Health Organization. Contact tracing in the context of COVID-19 (Interim Guidance), (https://apps.who.int/iris/rest/bitstreams/1329883/retrieve, accessed 27 November 2020).

Initiating Formal Methods with Z in the Philippines

Julian H. M. Rose

jrose.manila@gmail.com

ABSTRACT

Formal Methods of software engineering are an approach to software development that uses mathematics to describe and reason about software. We introduce the topic and focus on one notation in particular called Z; an international standard language for specifying software systems. We suggest ways for initiating formal methods and Z into both the educational curriculum and the workplace.

1 Introduction

Formal Methods of software engineering are an approach to software development that uses mathematics to describe and reason about software. They are mathematically rigorous techniques for the specification, development and verification of software. The use of formal methods is motivated by the expectation that, as in other engineering disciplines, performing mathematical analysis can contribute to the quality of a design.

1.1 Overview of formal methods

Numerous formal methods have been proposed since their introduction in the late 1960s through the pioneering work of Floyd [1] and Hoare [2]. In addition to predicate calculus and set theory, ground-breaking development in the 1930s in the work of Alan Turing, and the Lambda Calculus of Alonzo Church, have influenced the development of formal methods; and the 1950s gave rise to what are recognisably formal languages in the regular grammars of John Backus.

The Z language we cover in this article emerged during the late 1970s in the work of Jean-Raymond Abrial, and was rounded out at Oxford University during the 1980s, becoming industrial strength ready after the publication of Spivey's seminal text in 1989 [3]. A further decade of practical experience led to the international standardisation of Z in 2002 [4].

Formal methods like Z are mathematically-based languages engineered expressly for the capture of unambiguous requirements; their aim is precise communication between stakeholders. By counter-example in Figure 1, it is said a picture paints a thousand words.

Requirements evolve in adaptive projects. Requirements should be arranged into manageable pieces. Consistency across requirements needs to be maintained as changes and refactoring manifest. Traceability of requirements across design and into implementation is essential for acceptance testing and proving. Reasoning about formal requirements can be aided with proof assistants and CASE[1] tools.

In using Formal Methods usually more time will be devoted to developing requirements and in software acceptance testing, complemented by less time spent on maintenance and bug fixing [5, 6].

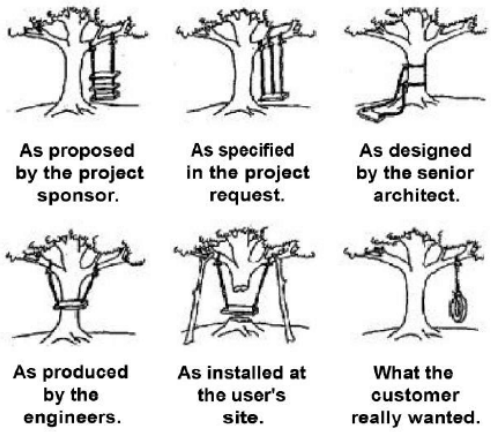


Figure 1. Pitfalls of poor requirements capture[2]

1.2 Overview of Z

The Z language was made an ISO standard language for software specification in 2002 [4]; and in 2007 technical corrections to the Z standard [7] were issued; otherwise no other changes to the 2002 standard have been adopted since and the Z language is stable.

Z is a formal language based on typed Zermelo-Fraenkel set theory combined with first-order predicate logic, characterised by a structural notation called schemas. Schemas together with the Z toolkit provide a high-level and concise way of writing good quality structured software specifications.

A good Z specification is a natural language document annotated with diagrams and Z paragraphs. This multi-content provides an understanding at multiple levels, helping to clarify the overall picture as well as essential detail. This article is representative of such a Z specification.

By analogy, to write and type-check Z is as to write and debug source code. Z specifications are fully typed such that they can be checked by CASE tools for syntax and consistency, not unlike program source code analysis. However, a Z specification is not (typically) executable in the sense that it cannot be run.

Z is used to model the subject domain and to describe *what* the software system is required to do through operational pre- and

[1] Computer-Aided Software Engineering

[2] The attribution to this well-known cartoon appears to be lost; maybe to Fred Brooks?

post-condition statements, rather than *how* it is to be done. A Z document commonly contains no detailed software or algorithm design, although specification refinement into design works towards that goal.

There are generally two classes of tools available to support Z: type checkers and proof assistants. Everyone who writes Z will use a type-checker; analogous to the way everyone who writes source code uses a compiler. Proof assistants are used in calculating theorems and testing specifications. They are considered a more specialist tool, perhaps even requiring a specialist role. Not everyone who writes Z will use a theorem prover.

The precision and well-formed nature of Z specifications makes the language applicable to high-value software development. By high-value we number safety-critical and high-risk business systems including life-support, telecommunications, aviation, as well as financial and business applications in the target set.

The use of Z is not restricted to software systems, or even computer-based systems, although that is the intended target. A Z specification can be used to help with a variety of tasks: most notably software development, but also hardware analysis, information systems design, project management, business operations analysis, amongst others.

To illustrate the use of Z we model a movie studio with films and actors in the remainder of this article.

2 Types

Every object in Z has a type. Remembering this and recognising what they are, along with type-checker tool support, will help to resolve many specification problems.

There is only one pre-defined type in standard Z, and that is the set of integers, $\mathbb{Z}$. (You will often see declarations of type $\mathbb{N}$, the natural numbers; $\mathbb{N}$ is actually a sub-type of $\mathbb{Z}$.) In particular there is no Boolean type in Z – it is good Z style to use relations instead. So when we start a new Z specification, one of the first things to be done is define our own types.

2.1 Sections

Near the beginning of a Z document it is usual (but optional) to identify a document using a section paragraph. (This is similar to providing a namespace for a source code file.) A section name groups declarations and definitions; making it easier to combine a section with other sub-system specifications.

section *FMZ_PH* parents *standard_toolkit*

There are two parts to a section paragraph: the section name, here *FMZ_PH*, and the comma-separated parents list, here *standard_toolkit*. Our declared section name is archived by the tools and is searchable. A document includes or inherits its parents' archives, providing it with access to all their definitions. The *standard_toolkit* section is built-in to the tools, and provides the high-level definitions (refer to [3, ch:4] [3]).

(**Tip**: When you start new Z specifications use a single column full-width page layout to make best use of space. This article uses two-column style as the PCJ preferred layout which is not conducive to Z text.)

The core (low-level) Z language is found in [3, ch:3]. The (high-level) mathematical toolkit (in [3, ch:4]) consists of definitions expressed in the core language which make writing specifications easier – a higher level of abstraction.

2.2 Types, objects and variables

Every object in Z has a type. That is the first thing to remember. There are five ways to define a type in Z: as a given type; a set type; a free type; a Cartesian tuple (including relations and functions); and as a Schema type. We will look at them all in turn, by example.

2.2.1 Given types

A given type is one in which there is no further structure; that is, it is a name for an unstructured object; unstructured means the specification takes no further interest in the type makeup – in reality objects may be fairly complex. This is the first way to define a type in Z. So, let's introduce our first given type:

[*Name*]

Here we define a new given type called *Name*. It is convention to capitalise types. A given type does not specify how a *Name* is constructed; whether there is a family name with a given name, or whether a title forms part of a name. Such information is just not of interest to the specification.

2.2.2 Individuals

In Z all individuals (or constants, instances, objects) are declared by name with their type. For example:

$$julia, block\_z, jinggoy, coming\_home, edgar, mia, bela, vbr, nash, the\_gift, tom, maledicto, cristine, untrue , lovi, malaya : Name$$

Here we declare a number of individual actors and movies of given type *Name*. In standard Z all individuals must be declared within scope, but not necessarily before use. The line down the side is the schema notation that signifies the individuals are declared as axiomatics.

2.2.3 Set types

In our movie studio model, we will use names for several different purposes: for actors, film titles and genres. The relationship between these different uses can be expressed using an axiomatic schema:

$$actors : \mathbb{P}\ Name$$
$$filmtitles : \mathbb{P}\ Name$$
$$genres : \mathbb{P}\ Name$$

$$actors \cap filmtitles \cap genres = \emptyset$$
$$actors \cup filmtitles \cup genres = Name$$

An axiomatic schema has global scope, meaning the content is visible throughout the document (or Z section). The line down the left-hand-side is the extent of the schema, and the line

[3] When we refer to Spivey [3] we should really refer to the standard [4], but Spivey remains a popular quick desktop reference.

across the middle separates the declarations above the line from the predicates or constraints below the line.

Above the line we have declared three *variables*: *actors*, *filmtitles* and *genres*. The '$\mathbb{P}$' means 'set of' (sometimes called *powerset* – see next tip), such that *actors* is of type *a set of Names*, and similarly for *filmtitles* and *genres*. Using a set is the second way of declaring a type in Z.

Below the line we have described the relationship between *actors*, *filmtitles*, *genres* and *Name*. The first predicate expresses that the sets of *actors*, *filmtitles* and *genres* are disjoint; that is no *Name* can appear in more than any one set. Through the second predicate, we say the three sets partition *Name*; that is, any *Name* instance must belong to one of the three sets and no other. We can depict the relationship within the axiomatic schema as a Venn diagram, in Figure 2.

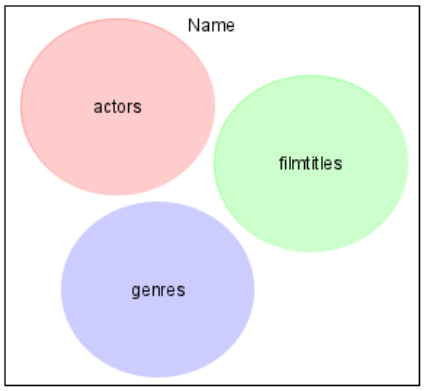


Figure 2. *Name* type and subsets

We can relate our individual actors with the set of actors by explicit set expression. This is given in another axiomatic schema:

$$actors = \{ julia, jinggoy, edgar, bela, nash, tom, cristine, lovi \}$$

Note that we do not need a declaration above the line; as the previous global declarations of *actors* and our individuals remain in global scope. Also notice the elements in a set are unordered, such that the set {*julia*, *jinggoy*} is the same as the set {*jinggoy*, *julia*}.

(**Tip**: Note *julia* is a *Name* but *{julia}* is a set of *Name*s (with one element). Whereas the elements of a type of set of *Names* are all its possible subsets.)

2.2.4 Free types

We can express the same structure and relationship between a type and its subsets using a Free type definition. For example, we can define awards for our movies:

$$Award ::= Gold \mid Silver \mid Bronze$$

Free types are the third way of defining a type in Z. Here the type *Award* is partitioned by the three disjoint constants *Gold*, *Silver* and *Bronze*; it shares the structure of Figure 2 (but not the labels). Free types are a convenience; their main advantage is that they make it easier to describe recursive structures such as lists and trees. (We provide a definition of lists in section 2.2.7.)

2.2.5 Variables and sub-types

We can use variables to stand for individuals. For example

$$x1 : \{ cristine, jinggoy, edgar \}$$

Here we declare a global variable, *x*, of type *{ cristine, jinggoy, edgar }*. More precisely *x* is of sub-type *{ cristine, jinggoy, edgar }* which has a type set of *Name* (because the type of the elements is *Name*). Therefore, *x* can hold any value of type *Name*, including for example *maledicto*. It would be better to declare *x* as a *Name* and restrict it with a predicate below the line.

2.2.6 Cartesian types

It is not meaningful to mix types within a set. For example, we cannot declare:

$$birthday : \{ nash, 1998 \}$$

since the types of *nash* and 1998 differ. To describe objects with more structure than given types or sets we can use Cartesian tuples; the fourth way of defining a type in Z. A tuple is a structured data type that may combine mixed types. For example:

$$birthday1 : Name \times Date$$
$$birthday1 = (nash, 1998)$$

in which the variable *birthday* is declared as a Cartesian type of *Name* cross *Date* , where *Date* is defined to be $\mathbb{N}$, the natural numbers:

$$Date == \mathbb{N}$$

We could just use $\mathbb{N}$ instead of defining *Date*, but our intentions are made clear by good use of named types. In a Cartesian type, the order of elements is significant, unlike a set; and we can refer to each element by its position in the tuple, so that:

$$birthday1 . 1 = nash$$
$$birthday1 . 2 = 1998$$

2.2.7 Mixing Cartesian and Free types

In addition to defining a Free type using constants, like *Award* above, Free types can be defined using constructors. To illustrate, we can define a list of actors as a recursive Free type structure together with a Cartesian type:

$$ListOfPeople ::= Person \langle\!\langle Name \rangle\!\rangle \mid PersonList0 \langle\!\langle Name \times ListOfPeople \rangle\!\rangle$$

A *ListOfPeople* is a Free type of two constructors, *Person* and *PersonList*. *Person* is constructed from a single *Name*; and *PersonList* is constructed from a pair of *Name cross ListOfPeople* and is thus a recursive definition. *ListOfPeople* allows us to declare a single variable that can take a list of *Name*s as a value. For example, if we want to shortlist potential actors for a new movie:

ShortList == ListOfPeople

candidates : *ShortList*

---

candidates =
*PersonList*0(*tom*,
*PersonList*0(*cristine*,
Person (*lovi*)))

which declares *candidates* as a global variable of type *ShortList* and populates it by an explicit expression. We use indentation to illustrate the structure of global variable *candidates* – this is not necessary in general.

The first *PersonList* comprises the *Name tom* and a second *PersonList* that comprises the *Name cristine* and a *Person* with *Name lovi*. This list data structure is illustrated in Figure 3.

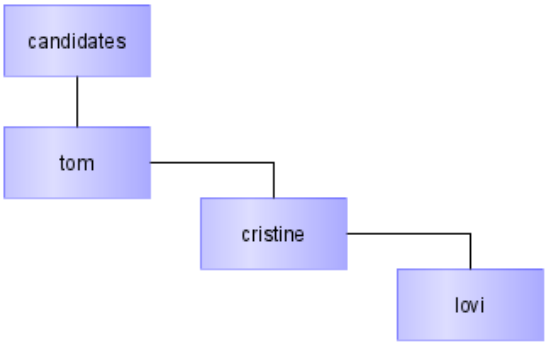


Figure 3. The list structure of candidates

2.2.8 Schema types

The first four ways of defining a type use the core language of Z. We can also use the schema language of Z to define types; the fifth and final way to define a type in Z. Suppose we want a *bio* for our candidate actors:

Bio

name : *Name*
birthdate : *Date*
salaryExpectation : ℕ

In the above schema, *Bio* is the only visible identifier at global scope; the horizontal lines at top and bottom delineate the scope of the schema content. *Bio* is a schema type that contains three elements of different types. It is similar to a Cartesian tuple, except that elements are identified by name rather than by position. The scope of all three elements falls within the local scope of the schema. We could define a global object, *julia_cv*, of type *Bio* and populate it thus:

*julia_cv*0 : *Bio*

---

*julia_cv*0 . *birthdate* = 1997
*julia_cv*0 . *name* = *julia*
*julia_cv*0 . *salaryExpectation* = 1000000

Or we can write out an explicit binding of the form:

*julia_cv*1 ==
⦇ *birthdate* == 1997,
name == *julia*,
salaryExpectation == *1000000* ⦈

The binding is not explicitly typed (such as *Bio*); the implicit type of a binding is derived from the types of its members, in this case *[birthdate*:ℤ, *name*:*Name, salaryExpectation*:ℤ*]*. This has the same type as *Bio* – remember it is the elements name and type, not their position, which characterise a schema type.

We can also write out a set comprehension:

julia_cv2 ==
{ *Bio* | *birthdate* = 1997 ∧
name = *julia* ∧
salaryExpectation = 1000000 }

In the above examples we have used indentation just for layout, which is not necessary in general.

2.2.9 Generic types

Suppose we want to keep a list of several different kinds of articles needed to make a film, such as camera equipment or costumes.

We can use a generic schema to specify a common list element for use with different kinds of articles:

Article [*X*]

uid : ℕ
category : *X*
allocated_to : *Name*

where *X* stands for a generic type which is instantiated on declaration.

But defining a list of Articles cannot be done using recursion within a schema. Except in the case of free type definitions, recursion is not allowed in Z. So we cannot define a schema containing a self-referential element (an element of its own type) *next : Article[X]* in the above.

The solution in Z is to define a Free type for the kinds of articles we want, for example:

Item ::= *Camera* | *Costume*

and to define a recursive list of Articles by Free type constructors of generic Articles (*Article[X]*) instantiated with the type of kinds of articles (*Item*):

ListOfItems ::=
ItemLast ⟪*Article[Item]*⟫ |
ItemList ⟪ *Article[Item]* × *ListOfItems*⟫

To add more kinds of items for listing all we have to do is add them to the *Item* free type. (We provide an example instantiation of *ListOfItems* in section 4.2.8.)

3 State

So far we have declared global variable of various types through axiomatic schemas. We will turn our attention now towards local declarations. In Z this is done using schemas which provide ways to organise a specification into manageable parts.

3.1 State schemas

In a film studio there are various people that are needed in order to shoot a scene for a film:

MoviePeople

$$stagehands : \mathbb{P}\ Name$$
$$filmcrew : \mathbb{P}\ Name$$
$$directors : \mathbb{P}\ Name$$
$$actors : \mathbb{P}\ Name$$

$$stagehands \cap filmcrew \cap directors = \emptyset$$
$$stagehands \cap filmcrew \cap actors = \emptyset$$

In the above schema, *MoviePeople* is the only visible identifier at global scope; the horizontal lines at top and bottom delineate the scope of the schema. Through the top and bottom delineation, schemas achieve local scope. Identifiers *stagehands*, *filmcrew*, *director* and *actors* are only visible locally, within the schema itself. In particular note that *MoviePeople* . *actors* is local and does not conflict with the global *actors* definition in section 2.2.3 above. The selection of either *actors* variable is determined by the scope of its declaration; within the schema, reference to *actors* uses the local declaration.

The difference between a schema type and a state schema is not a definite line. In *Bio*, we identify a schema type as a structure containing multiple elements which may be of mixed type; whereas, in *MoviePeople*, we identify a state schema as a structure containing elements which are variables of a *set of some type* (the types may also be mixed). However, it is quite usual to have a schema type with set type elements, and with constraints below the line.

Below the line in *MoviePeople* we constrain people from having more than one role on a film set; except that *actors* may also be *directors*.

3.2 Constraints

In addition to *MoviePeople*, another part of the modelled state is the produced films.

$$Result ::= Accept \mid Reject$$
$$Scene == \mathbb{N}$$
$$Take == \mathbb{N}$$

We define a global Free type *Result* as the constants *Accept* or *Reject*. We declare that *Scene* and *Take* are natural numbers, $\mathbb{N}$. We could just use $\mathbb{N}$ instead, but giving names to our types makes them more meaningful. A film comprises a number of *FilmSegments*:

FilmSegment

$$action : (Scene \times Take) \pfun Result$$

$$\forall s : Scene \bullet$$
$$\exists t1, t2 : Take \bullet$$
$$(action\ (s, t1) = Accept \land action\ (s, t2) = Accept) \Rightarrow t1 = t2$$

3.3 Functions

FilmSegment declares one local variable, *action*, which is a function (a form of Cartesian tuple type). The left hand side, or argument, or domain of the function is a Cartesian pair, *Scene* × *Take*; and the right hand side, or result, or range is of type *Result*. More specifically *action* is a *partial function* indicated by the vertical bar on the arrow; a partial function is one in which only some of its domain are mapped to a result, illustrated in another Venn diagram in Figure 4.

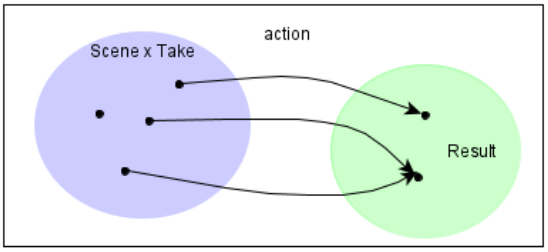


Figure 4. A snapshot of partial function *action*

Figure 4 illustrates four points in the domain of *action*, representing pairs of *Scene* × *Take*, and two points in the range of *action*, representing the possible values of *Result*. Only three of the four points in the domain are mapped to points in the range in this instance of *action*. The type of *action* is *Scene* × *Take* × *Result*, a Cartesian triple.

A function in Z is a data structure, not to be confused with an Operation in which a state transition is described. We cover operations below in section 4.

3.4 Quantifiers

Below the line in *FilmSegment* is a constraint on the *action* function. The expression $\forall s : Scene$ reads for all *s* of type *Scene*, and introduces a variable, *s*, that stands for any value taken from the set *Scene*. $\exists t1, t2 : Take$ reads there exists *t1*, *t2* of type *Take*, and introduces two variables, *t1* and *t2*, that each stand for a possible value taken from the set *Take*. (Refer to [3, section 3.7].)

Informally the constraint below the line says that for any *scene* *s* if there exists a take, *t1*, paired with *s* which results in *Accept*, and there exists a *take*, *t2*, paired with *s* that results in *Accept*, then it must be the case that *t1* = *t2*. In other words, there is only one *take* paired with any scene which results in *Accept*. This is a classically Z way of expressing uniqueness in a constraint. (There may be any number of *take*s paired with any one *scene* which result in *Reject*.)

3.5 Sequences

There is a second state schema we want to add into our model state space, and that is a *Film*:

Film

$$movie : \text{seq } FilmSegment$$

$$\forall\ s : Scene;\ t : Take;\ r : Result;\ f : FilmSegment \mid f \in ran\ movie \land ((s, t), r) \in f.action \bullet r = Accept$$

The schema *Film* introduces a local variable, *movie*, of type sequence of *FilmSegment*. A sequence is distinguished from a set in that its elements are ordered, in sequence. The set $\{ a, b, c \} = \{ c, b, a \}$; whereas the sequence $\langle a, b, c \rangle \neq \langle c, b, a \rangle$.

The constraint below the line is read for all *s* of type *Scene*, *t* of type *Take*, *r* of type *Result*, *f* of type *FilmSegment*, filter, *f* is an element of the range of *movie*, and the triple *(s, t, r)* is an element of *f . action*, such that *r* equals *Accept*. Informally it means a *Film* contains no out-takes.

3.6 State Space

Typically, a Z specification includes both static and dynamic aspects of the system being modelled. Static aspects include the states a system can occupy and the invariant relationships that hold across all possible states of the system. This is called the State Space (of a model).

Our state space is the combination of our state schemas, *Film* and *MoviePeople*; brought together into a single schema named *MovingPicture*

MovingPicture

$$staff : MoviePeople$$
$$picture : Film$$
$$appointed : \mathbb{P}\ Studio$$
$$works\_on : MoviePeople \leftrightarrow Film$$

$$staff \in dom\ works\_on$$
$$picture \in ran\ works\_on$$

which also introduces a local variable *appointed* of type set of given type Studio

[*Studio*]

to record which *Studios* are allocated to the *MovingPicture*, and a relation, *works_on*, to relate the combined state schemas. The domain (or left-hand side) of *works_on* is bound to *staff*, and the co-domain (or right-hand side) of *works_on* is bound to *picture*. Thus our state space records who works on which pictures.

4 State Transitions

Typically, a Z specification includes dynamic aspects of the system being modelled. Dynamic aspects are the operations a system performs and invariant relationships that hold across those operations.

4.1 Transition semantics

The approach to state transition in Z is not that of a state machine - or any kind of abstract machine - there is no syntax or convention for stringing together operations in sequence, such as "do operation *a* then do operation *b*". To put it another way, there is no time in Z[4].

(**Tip**: Some try to sequence operations by defining a free type of state transitions; but like defining a free type of Booleans, this is considered poor Z style and should be avoided. A good approach, though one for design rather than specification, is to model a state machine in Z with operations over the domain model state space.)

The usual semantics of operations in Z derives from its logic foundation: if the operation preconditions are met, then the outcome is known; whereas if the preconditions are not met, then the outcome is unknown. In either case the operation transition may take place.

That is different to firing condition semantics that some assume: if the preconditions are met then the operation is enabled (can fire); whereas if the operation preconditions are not met then the operation is disabled (cannot fire). This assumption is not wrong, but must be made expressly and used consistently throughout a project (not just the containing document or section)[5].

4.2 Operations

An operation in Z makes use of schemas, such that the operations of a system are organised into manageable pieces. An operation describes what is to be done but not how it is done. We can say, for example, that a list is to be sorted, but we do not say which sorting algorithm shall be used.

4.2.1 Pre- and post-conventions

An operation in Z is described using pre- and post-conditions. Suppose we hire an actor for a movie; this can be (incompletely) expressed:

*Hire*0

$$MovingPicture$$
$$MovingPicture'$$
$$actor? : Name$$

$$staff'.actors = staff.actors \cup \{ actor? \}$$

(*Hire*0 is incomplete and we will address this shortly.) Notice there are two state schema inclusions, one decorated (with a ') and the other undecorated. By convention the undecorated schema, *MovingPicture*, stands for the pre-state and the decorated schema, *MovingPicture '*, stands for the post-state. The pre-state is the state of our system prior to transition of the operation *Hire0*. The post-state is the state of our system after transition of the operation *Hire0.*

Notice also the local variable, *actor?*. By convention in Z, a variable decorated with a '*?*'is an input variable to the operation schema (and '*!*' is an output variable decoration). Recall the

[4] Mostly true with minor exceptions. See section 4.2.9 for example.

[5] There are cautionary tales of mixing units, like centimeters and inches.

scope of a schema starts and finishes with the top and bottom line, so the declarations are all local in scope.

The *Hire*0 operation is specified by the predicate, or invariant, 'below the line' in the middle. Informally this says the post-state equals the pre-state union the value of *actor?*; in other words, the input actor is added to our set of actors, or 'hired'.

4.2.2 More complete operation

As mentioned, Hire0 is incomplete. Let's look at a complete operation, Hire1:

*Hire*1

$$
\begin{aligned}
&MovingPicture \\
&MovingPicture\,' \\
&actor? : Name \\
&\hline
&staff'.actors = staff.actors \cup \{ actor? \} \\
&staff'.stagehands = staff.stagehands \\
&staff'.filmcrew = staff.filmcrew \\
&staff'.directors = staff.directors \\
&picture' = picture \\
&appointed' = appointed \\
&works\_on' = works\_on
\end{aligned}
$$

A complete operation must explicitly state that state not explicitly affected by the operation is not affected by the operation; in other words that nothing else changes, or that there are no side-effects of this operation. (This is not a peculiarity of Z, but is a truth in languages based on first-order logic.)

(**Tip**: we cannot write $staff'.actors = staff.actors \cup \{ actor? \}$ together with $staff' = staff$, because that is inconsistent, like writing 1 = 0.)

4.2.3 More operation conventions

The delta 'Δ' convention is a shorthand convention in Z for including both the pre- and post-state of a state schema:

*Hire*2

$$
\begin{aligned}
&\Delta MovingPicture \\
&actor? : Name \\
&\hline
&staff'.actors = staff.actors \cup \{ actor? \} \\
&staff'.stagehands = staff.stagehands \\
&staff'.filmcrew = staff.filmcrew \\
&staff'.directors = staff.directors \\
&picture' = picture \\
&appointed' = appointed \\
&works\_on' = works\_on
\end{aligned}
$$

We still have to write out all the post-states however.

4.2.4 Slightly better operation

Now, we haven't considered whether *actor?* is already hired. We can write a slightly better relation between pre- and post-members of staff shown in *Hire3*:

*Hire*3

$$
\begin{aligned}
&\Delta MovingPicture \\
&actor? : Name \\
&\hline
&staff'.actors = \\
&\quad (staff.actors \setminus \{ actor? \}) \cup \{ actor? \} \\
&staff'.stagehands = staff.stagehands \\
&staff'.filmcrew = staff.filmcrew \\
&staff'.directors = staff.directors \\
&picture' = picture \\
&appointed' = appointed \\
&works\_on' = works\_on
\end{aligned}
$$

Hire3 ensures that *actor?* is removed from the pre-state *staff*.*actors* set, if present, before being added to the post-state *staff'*.*actors* set. *Hire3* is in effect two operations in one (done at the same time).

4.2.5 Error conditions

Best practice is to provide pre-condition checks for operations. Here, we can first check whether *actor?* is already hired:

*Hire*4

$$
\begin{aligned}
&\Delta MovingPicture \\
&actor? : Name \\
&\hline
&actor? \notin staff.actors \\
&staff'.actors = staff.actors \cup \{ actor? \} \\
&staff'.stagehands = staff.stagehands \\
&staff'.filmcrew = staff.filmcrew \\
&staff'.directors = staff.directors \\
&picture' = picture \\
&appointed' = appointed \\
&works\_on' = works\_on
\end{aligned}
$$

In Hire4 we add a pre-condition that *actor?* is not an element of the set *staff*.*actors*. Now, bearing in mind the usual semantics of Z tells us that an operation can always transition, that when its' preconditions are not met the outcome is not defined, the recommended solution is for operations to output results.

Success

$$
\begin{aligned}
&out! : Result \\
&\hline
&out! = Accept
\end{aligned}
$$

Fail

$$
\begin{aligned}
&out! : Result \\
&\hline
&out! = Reject
\end{aligned}
$$

And we can use the schema calculus of Z to combine schemas, defining our final *Hire* operation:

$$Hire == (Hire4 \land Success) \lor ((\neg Hire4) \land Fail)$$

With *Hire*, informally, if the pre-conditions and post-conditions of *Hire4* are true *Success* will output the value *Accept*; or if they are not true *Fail* will output the value *Reject*. (The expression ¬*Hire4* negates the predicate below the line.)

4.2.6 Schema calculus

Z encompasses two logic languages: predicate calculus which is found within schemas in the constraints and relationships between variables 'below the line', and the schema calculus which operates on schemas enabling them to be combined into new structures.

Using the schema calculus, a Z specification can be arranged into manageable pieces, or modules, or building blocks; and those pieces may then be combined to make larger specification structures. Thus, we can say a Z specification is a structured specification, in which small mathematical pieces can be developed in isolation and these can be (re-) used to construct larger pieces.

Writing out the left-hand side of *Hire* gives:

Hire_lhs

$$\Delta MovingPicture$$
$$actor? : Name$$
$$out! : Result$$

$$(actor? \notin staff\ .\ actors$$
$$staff'.\ actors = staff\ .\ actors \cup \{ actor? \}$$
$$staff'.\ stagehands = staff.\ stagehands$$
$$staff'.\ filmcrew = staff.\ filmcrew$$
$$staff'.\ directors = staff.\ directors$$
$$picture' = picture$$
$$appointed' = appointed$$
$$works\_on' = works\_on)$$
$$\land$$
$$(\ out! = Accept\ )$$

in which the declarations from (*Hire*4 ∧ *Success*) are merged above the line, and predicates from the combined schemas are merged under conjunction below the line.

Writing out *Hire_rhs* is similar. And the whole schema Hire can be written out with the predicates of *Hire_lhs* and *Hire_rhs* merged under disjunction below the line, with the common set of declarations above the line.

4.2.7 Read-only operations

There is another shorthand convention in Z for including both the pre- and post-state of a state schema and this is the *xi* 'Ξ' convention. This convention is used to describe read-only operations. For example, we can test whether and actor is hired:

*IsHired*0

$$\Xi MovingPicture$$
$$actor? : Name$$

$$actor? \notin staff\ .\ actors$$

The 'Ξ' convention states that the post-state equals the pre-state. Since in using Ξ there is no change to the state, we may ask what purpose the 'Ξ' convention serves; why not just write

*IsHired*1

$$MovingPicture$$
$$actor? : Name$$

$$actor? \notin staff\ .\ actors$$

without the decoration? By using the 'Ξ' convention we make our intention clear that the operation is a read-only operation and the state space is not changed. Any future modifications to the operation need to respect that intent; if in the future some change in state does become necessary then a new operation using the 'Δ' convention should be developed instead, leaving the read-only operation intact.

4.2.8 Operations on generic data types

We can define an operation that works entirely on input variables rather than the state space. For example, with reference to section 2.2.9 above, we can search the UIDs of all items of a particular kind in a list:

SearchUidsInList

$$getUids : ListOfItems \rightarrow \mathbb{P}\ \mathbb{N}$$
$$l? : ListOfItems$$
$$cat? : Item$$
$$out! : \mathbb{P}\ \mathbb{N}$$

$$\forall a : Article \bullet$$
$$(\ a.category = cat?\ \land$$
$$getUids(\ ItemLast(\ a\ )) = \{\ a.uid\ \}\ )$$
$$\land$$
$$getUids(\ ItemList(\ a, l?\ )) =$$
$$getUids(\ ItemLast(\ a\ )) \cup getUids(\ l?)$$
$$out! \in ran\ getUids$$

In *SearchUidsInList* we define a local function, *getUids*, which provides the data structure needed for the operation; two input variables, *l?*, the *ListOfItems* to be searched, and *cat?*, the kind of *Item* to search for within the list; and one output variable, *out!*, which is the resulting set of UID numbers matching the search criteria. *getUids* is in the style of a recursive function. Understanding what the search is required to do is aided by considering the structure of the Free type *ListOfItems* defined in section 2.2.9 above (refer to [3, section 3.10] and also [6, section 10.4]):

i/ *ItemLast* is a function (an injection) from *Article* to *ListOfItems*:

$ItemLast : Article[Item] \rightarrowtail ListOfItems$

it is enough to match the *Item* kind and extract the UID.

ii/ *ItemList* is likewise an injection, from a pair *Article cross ListOfItems* to *ListOfItems*:

$ItemList : (Article[Item] \times ListOfItems) \rightarrowtail ListOfItems$

and each element of the pair is examined separately. The *Item* kind element is extracted to create a new *ItemLast* element, and this is used in a recursive application to *i* above[6]. The tail of the list element is examined in a recursive style. (The order in which this pair of elements is tested is not specified – conjunction doesn't imply the left-hand is evaluated before the right-hand).

(**Tip**: If in designing a Z operation you find yourself trying to 'write program code' then stop because you are going wrong. An operation in Z describes what is to be done; not how the algorithm works. In other words we don't design the code to do a search, we write 'search for these things'.)

4.2.9 Initialisation operations

Returning to our state space, we need to be able to initialise the state to establish a starting point. (This is one case, by convention, where some notion of time is present in a standard Z specification.) This is done using an initialisation operation.

Init

$\Delta MovingPicture$

$staff'.stagehands = \{\}$
$staff'.filmcrew = \{\}$
$staff'.directors = \{\}$
$staff'.actors = \{\}$
$picture'.movie = \langle\rangle$
$appointed' = \{\}$
$works\_on' = \{\}$

With the *Init* operation, all post-state variables are initialised to the value 'empty set'. (Note it is not necessary to initialise *works_on* as by definition that function is bound to *staff* and *movie* which are initialised; but we do anyway to err on the side of safety.) No tests are made on pre-state variables – equivalent to having a pre-state value *true*. This is because we want to be able to initialise our state space into a known outcome no matter what present state exists.

5 Promotion

So far we have only described individual movies. Suppose our film studio wants to maintain a catalogue of its *MovingPicture* productions. (This information might be held in a database with one entry for each *MovingPicture*.) Then there exists a relationship between the catalogue state and the state of each individual film, and this relationship can be used to link operational changes to the (global) catalogue with changes to individual (local) films.

5.1 A film catalogue

We introduce the schema *Catalogue* to contain our global state space:

Catalogue

$directory : filmtitles \rightarrow MovingPicture$
$classification : genres \nrightarrow MovingPicture$
$vintage : Date \nrightarrow MovingPicture$
$assigned : Studio \nrightarrow MovingPicture$

Catalogue is related to individual film local state space through the functions *directory*, *classification*, *vintage* and *assigned*.

directory is a total function meaning every element of its domain, *filmtitles*, is assigned a value – there exists no *filmtitle* that is not in the *directory*; whereas *classification*, *vintage* and *assigned* are all partial functions, so that a subset of their domains (but possibly every element) is assigned a value. For example, it is possible to have a *genre*, say *biopic*, for which there are no films in the catalogue; or some years in which no films were made.

5.2 Global operation

A movie may need to have a studio *assigned* to shoot a film sequence, releasing it after the shoot. The global operation to assign a studio is captured in *BookStudioG*:

BookStudioG

$\Delta Catalogue$
$\Delta MovingPicture$
$s? : Studio$
$p? : MovingPicture$

$s? \notin dom\ assigned$
$s? \notin appointed$
$\wedge$
$assigned'(s?) = p?$
$appointed' = appointed \cup \{s?\}$
$\wedge$
$\{s?\} \lhd assigned' = \{s?\} \lhd assigned$
$directory' = directory$
$classification' = classification$
$vintage' = vintage$
$staff' = staff$
$picture' = picture$
$works\_on' = works\_on$

The pre-condition for *BookStudio* is that the input studio *s?* is neither in the global state variable *assigned* nor the local state variable *appointed*. The operation assigns the input *s?* accordingly and asserts no other changes are to be made. Of particular interest is the line

$\{s?\} \lhd assigned' = \{s?\} \lhd assigned$

[6] We don't write $\forall a:Article \mid a.category = cat?$ before the scoping • otherwise *ItemLists* containing non-matching *Article* kinds would be excluded from the search.

which uses domain restriction ($\lhd$) to specify that all elements of the partial function *assigned* except that mapping *s?* remain unchanged; in other words, all other assigned studios remain untouched by the operation.

5.3 Refactoring the global operation

As schemas are combined to construct larger specification items from smaller pieces, so factoring allows global and local operations to simplify the global specification structure. We may restructure or *refactor* a global operation into a local operation and a mixed operation, the latter expressing the relationship between local and global state space.

Such specification restructuring provides a useful separation of concerns; the two factors (local operation and mixed operation) may be specified and analysed independently (refer to [6, section 13]).

5.3.1 Local operation

BookStudioL

$\Delta MovingPicture$
$s? : Studio$

$s? \notin appointed$
$\wedge$
$appointed' = appointed \cup \{s?\}$
$\wedge$
$staff' = staff$
$picture' = picture$
$works\_on' = works\_on$

The operation *BookStudioL* works on the local state space. It records that the input *Studio s?* is added to *appointed*, the function which records the set of studios allocated to the *MovingPicture*. As usual we clarify that all other delta state remains unchanged by this operation.

5.3.2 Mixed operation

The mixed operation *BookStudioP* works on the global state space. It records that input *Studio s?* is added to *assigned*, the function which records the set of studios allocated to the set of *MovingPicture*s.

BookStudioP

$\Delta Catalogue$
$MovingPicture'$
$s? : Studio$

$s? \notin dom\ assigned$
$\wedge$
$assigned' =$
$\quad assigned \oplus \{ s? \mapsto \theta\ MovingPicture' \}$
$\wedge$
$directory' = directory$
$classification' = classification$
$vintage' = vintage$

As usual we clarify that all other delta state remains unchanged by this operation. Of particular note is the line

$assigned' =$
$\quad assigned \oplus \{ s? \mapsto \theta\ MovingPicture' \}$

which uses relational override ($\oplus$, a form of union) to state that *assigned'* includes the mapping from the input *s?* to a *binding* (single instance) of *MovingPicture'*. Exactly which binding is made known through the schema calculus in the promoted operation.

5.3.3 Promoted operation

The promoted operation, that re-structured or *refactored* operation comprising a local and a mixed operation, is defined using the schema calculus:

$BookStudio ==$
$\quad BookStudioL \wedge BookStudioP$

6 Proof

We use proof and calculus to reason about specifications in Z. One such evaluation is to check the precondition of each operation is complete.

6.1 Calculate the pre-condition

To check the pre-condition of an operation, in general:

$Operation == [\ \Delta Declaration \mid Predicate\ ]$

we first divide $\Delta Declaration$ into:

i. *Before*, containing only the un-primed state components and input variables;
ii. *After*, containing only the primed state components and outputs, giving:

$pre\text{-}Operation ==$
$\quad [\ Before \mid \exists\ After \bullet Predicate\ ]$

For example, the precondition for the operation *Hire_lhs* from section 4.2.6 by applying steps *i* and *ii* above is:

pre_Hire_lhs_ii

$MovingPicture$
$actor?: Name$

$\exists\ out! : Result;$
$\quad MovingPicture' \bullet$
$\quad\quad (actor? \notin staff\ .\ actors$
$\quad\quad staff'\ .\ actors =$
$\quad\quad\quad staff\ .\ actors \cup \{\ actor?\}$
$\quad\quad staff'\ .\ stagehands = staff\ .\ stagehands$
$\quad\quad staff'\ .\ filmcrew = staff\ .\ filmcrew$
$\quad\quad staff'\ .\ directors = staff\ .\ directors$
$\quad\quad picture' = picture$
$\quad\quad appointed' = appointed$
$\quad\quad works\_on' = works\_on)$
$\quad\quad \wedge$
$\quad\quad (\ out! = Accept\ )$

iii. Next, expand schemas in *After* :

pre_Hire_lhs_iii

$$
\begin{array}{|l}
MovingPicture \\
actor?: Name \\
\hline
\exists\, out! : Result; \\
\quad staff\_\_actors' : \mathbb{P}\ Name; \\
\quad staff\_\_stagehands' : \mathbb{P}\ Name; \\
\quad staff\_\_filmcrew' : \mathbb{P}\ Name; \\
\quad staff\_\_directors' : \mathbb{P}\ Name; \\
\quad picture' : Film; \\
\quad appointed' : \mathbb{P}\ Studio; \\
\quad works\_on' : MoviePeople \leftrightarrow Film \bullet \\
\qquad (actor? \notin staff\ .\ actors \\
\qquad staff\_\_actors' = \\
\qquad\quad staff\ .\ actors \cup \{ actor?\} \\
\qquad staff\_\_stagehands' = staff\ .\ stagehands \\
\qquad staff\_\_filmcrew' = staff\ .\ filmcrew \\
\qquad staff\_\_directors' = staff\ .\ directors \\
\qquad picture' = picture \\
\qquad appointed' = appointed \\
\qquad works\_on' = works\_on) \\
\qquad \wedge \\
\qquad (\ out! = Accept\ ) \\
\hline
\end{array}
$$

(Note we had to rename the members of *MovingPicture'* as it is neither legal Z syntax nor meaningful to declare variables through schema dot membership.)

iv. Lastly, use the equations in *Predicate* to eliminate terms with post-state and output variables:

- eliminate *staff__actors'* by *staff .actors* $\cup \{actor?\}$;
- eliminate *staff__stagehands'* by *staff .stagehands*; similar for *staff__filmcrew'* and *staff__directors'*;
- eliminate *picture', appointed'* and *works_on'* by *picture, appointed* and *works_on*, giving:

pre_Hire_lhs_iv

$$
\begin{array}{|l}
MovingPicture \\
actor?: Name \\
\hline
\exists\, out! : Result \bullet \\
\qquad (actor? \notin staff\ .\ actors) \\
\qquad \wedge \\
\qquad (\ out! = Accept\ ) \\
\hline
\end{array}
$$

We cannot eliminate *out!* as there is no equation involving it in *Predicate*. Thus *pre_Hire_lhs_iv* is the calculated pre-condition for operation *Hire_lhs*.

In this case we can see nothing has been missed from the operation specification for *Hire_lhs* – there is nothing we have forgotten to include in the pre-condition. (Refer to [6] for a detailed description of pre-condition calculation.)

6.2 Wider proof and formal methods

Now, we started this article by stating that formal methods are an approach to software development that uses mathematics to describe *and reason* about software.

Calculating the pre-condition of operations is just one application of proofs in Z. We have barely touched on the wider topic of proofs and calculus for Z; indeed proofs probably account for half of the subject matter of Z, especially if taking refinement of Z specifications into account. Refinement of (subject domain) specifications is the process of walking them towards (computing domain) implementation in a stepwise manner, usually through a series of calculations or *refactoring*.

However, refinement does not expressly address the matter of design, and assumes the structure of a design or implementation is directly related to the structure of the specification, such that one can be calculated from the other. It is not that Z or refinement prevents design; it just is not given much consideration in articles and text books.

7 FMs in the Philippines

7.1 Surveys

Use of formal methods is motivated by the expectation that performing mathematical analysis can contribute to the quality of a design. However, there have been relatively few quantitative studies that demonstrate positive or negative support for that assertion. Quantitative analysis of formal methods in the Philippines would be of great benefit. Surveys of use, experience, need or interest would help to steer research agendas and educational programming tailored to local needs.

7.2 Educational Programming

It is interesting to note Asian countries which include formal methods in the educational curriculum, notably India and China, enjoy a strong presence in the global software industry. Though we do not claim a direct correlation between formal methods in the curriculum and a healthy software industry here, we do have formal methods education in our sights in the Philippines.

If people are taught how to program in Java, then they can program in Java; but if they are taught programming fundamentals, then they can become skilled at programming in different languages. Similarly, if people are taught software engineering fundamentals including formal methods, then they could be more able to take on significant (architectural) roles in the specification and design of software systems.

Educational programmes for formal methods are most-often developed for undergraduate and graduate student degree levels. The subject can be delivered as modules over several undergraduate grade years, possibly as an elective subject, or as a self-contained course, or on a taught Master's degree.

7.3 Technology Transfer

Professional top-up programmes in formal methods can be assembled for working engineers and managers. These programmes would help to align the Philippines with best practice in Asia, and help extend services offered by Philippine companies.

7.3.1 UML and Z

Introducing Z into the workplace and the educational curriculum might be eased by combining Z with non-formal methods. The aim is to apply more user-friendly notations,

diagrams and methods (ways of doing things) which are complemented by the formal language. Deriving a UML design from a Z specification, rather than doing formal refinement and proofs, might make for an easier initiation of (semi-)formal methods.

(**Tip**: There is an object-oriented flavour of Z[7] called Object-Z which is in print [8]; although it has never been made a standard language like Z. Object-Z is supported by the CZT tools.)

Appendix. How to start with Z

Just reading an article or book on Z, or sitting in on slide presentations, is not enough. Like programming, to grasp Z people must do Z (college professors take note).

Anyone can start writing and type-checking Z by downloading the free ZWordTools, a plug-in for Microsoft Word; or the CZT plug-in for JEdit; or download the standalone java CZT tools. (CZT is the Community Z Tools, an Open Source project.)

(**Tip**: on installing ZWordTools de-select the option 'load ZWordTools every time Word starts'; it takes seconds to execute the start-up script, which you would want to avoid most times you use Word.)

As a first exercise the film studio example in this article can be entered and type-checked to remove typos and familiarise with the tools and the Z language. The film studio model could be extended with popularity features such as likes or ratings.

Writing specifications with Z is a social endeavour. A group of friends, classmates or work colleagues can meet (online) to try out some ideas and share mutual help. Z can be done in a coffee-shop or the mall, as well as the classroom and the lab, or working from home.

There are several well-known Z texts that are available online [3, 7, 9, 10], (including the Z standard for reference [4]). You can download these and study them. Other recommended references (not online) include [11] for its pedagogic approach and [12] for its simplicity. The latter was used by this author when teaching a University course on Z.

And there are many downloadable articles concerning Z available online at websites including SemanticScholar.org and ResearchGate.net. So you can find topics that interest you specifically for further research.

Acknowledgements

The Z language is very much the work of others and the author, a private researcher, is acting as an ulat ng manunulat and occasional commentator. We thank the editors for bringing this article into being and the anonymous referee, in particular for pointing out an error in Figure 3.

This document was prepared using Microsoft Word 2007 with the plug-in ZWordTools Release 3.3.0.1 configured for CZT type-checking with option "Allow use before declaration". We have over-used brackets in our Z descriptions to make clear our intentions and not solely rely on Z operator precedence (following good programming practice).

[7] There exist several different object-oriented Z proposals; Object-Z is complete.

References

[1] R.W. Floyd. (1967). *Assigning meanings to programs*. Proceed. Symposia in Applied Math., vol. 19.

[2] C.A.R. Hoare. (1969). *An axiomatic basis for computer programming*. Comm. Of the ACM, vol 12(10).

[3] J.M. Spivey. (1992). *The Z Reference Manual 2nd edition*. Prentice Hall International (UK).

[4] *ISO IEC 13568: Information technology — Z formal specification notation — Syntax, type system and semantics*, July 2002.

[5] A.J. Hall. (1990). *Seven myths of formal methods.* IEEE Software 7(5), pp 11-19.

[6] J. Woodcock, J. Davies. (1996). *Using Z: Specification, Refinement and Proof*. Prentice Hall International.

[7] *ISO IEC 13568:2002/COR 1:2007 – Information technology — Z formal specification notation — Syntax, type system and semantics* — Technical Corrigendum 1, July 2007.

[8] G. Smith. (2000). *The Object-Z specification language*, Springer.

[9] I.J. Hayes. (1992). *Specification Case Studies*, Prentice-Hall.

[10] J.P. Bowen (ed). (2003). *Formal Specification and Documentation using Z: a case study approach*, Intl Thomson Publishing.

[11] B. Potter, J. Sinclair, and D. Till. (1996). *An introduction to formal specification and Z* (2nd edition), Prentice Hall.

[12] D. Lightfoot. (2000). *Formal specification using Z (2nd edition)*, Macmillan.

Effect of Parameters and Clustering Algorithms on Interaction-Based Community Detection in Twitter

Ryan Austin Fernandez, Clarisse Felicia M. Poblete, Marc Dominic San Pedro
Johansson E. Tan, Charibeth K. Cheng
De La Salle University - Manila
Manila, Metro Manila
{ryan.fernandez,clarisse_poblete,marc_sanpedro,johansson_tan,charibeth.cheng}@dlsu.edu.ph

ABSTRACT

In a social network, communities are groups of users that are more similar to each other, based on certain well-defined interactions, than they are to users outside their communities. Community detection is defined as the extraction of the communities in these networks. This study compares different combinations of clustering algorithms, similarity measures, and social network features, and analyzes the differences of the produced communities. Multiple iterations of community detection were performed on data gathered from Twitter. Some selected features produced communities that reflect interactions among the users in the network, while others produced communities that reflect common interests. Algorithms with *simulated annealing* produced more communities (100-900) than those without (at most 80).

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**.

KEYWORDS

community detection, clustering, social networks

ACM Reference Format:
Ryan Austin Fernandez, Clarisse Felicia M. Poblete, Marc Dominic San Pedro and Johansson E. Tan, Charibeth K. Cheng. 2020. Effect of Parameters and Clustering Algorithms on Interaction-Based Community Detection in Twitter. In *Workshop On Computation: Theory And Practice, November 21, 2020, Online Conference.* ACM, New York, NY, USA, 10 pages. https://doi.org/

1 INTRODUCTION

Social media has become more prevalent in recent years. People participate in *microblogging*, where they share their thoughts, status, and opinions in short posts on a social network such as Twitter [10]. Posts are limited to two-hundred and eighty characters. These social media platforms are an opportunity to mine sentiments and detect patterns in the social network. One such pattern that can be found in social networks is the concept of a *community*.

WCTP 2020, November 21, 2020, Online Conference

ACM ISBN …$
https://doi.org/

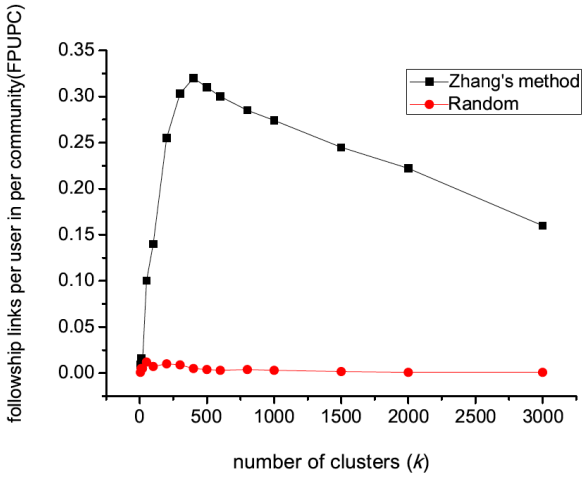


Figure 1: Results of Zhang's community detection as compared to random clustering [21]

Communities are defined as groups of users that are more similar to each other than they are to users outside their communities [6]. Community detection is necessary because it shows the interaction of multiple users with respect to specific features such as user following, post topics, and user mentions. Numerous studies on community detection have already been done including [12, 19, 21].

The problem in community detection is the abundance of *hyperparameters* when it comes to setting up the community detection system. Among these parameters are the features that will be used to associate different users to one another, the actual detection algorithm to be used to extract the communities, and the evaluation metric to be used to determine the quality of the produced communities [21]. Naively or randomly clustering together users has also been found to be less effective than an actual algorithm, as shown in Figure 1.

Most of the pre-existing works by [5, 13, 19] either had preselected features and algorithms to work with or specific features in mind before implementing their community detection system. In this study, we investigate the effects of using different combinations of detection algorithms, similarity parameters, and evaluation metrics on the extracted communities in terms of trends and structure.

We begin by discussing the various detection algorithms, similarity parameters, and evaluation metrics that were implemented in this study. We then describe the dataset used in our testing of the

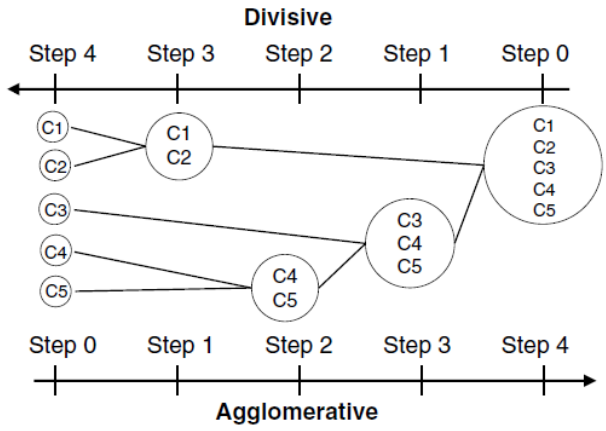


Figure 2: Divisive versus Agglomerative Hierarchical Clustering from [2]

algorithms. Next, we explain how our experiments were performed. Finally, we present and discuss the results of the experiments.

2 PRELIMINARIES

This study deals with the analysis of the communities formed by different combinations of community detection algorithms and similarity measures. Each of the community detection algorithms and similarity measures is explained in this section, as well as the community evaluation metrics.

2.1 Community Detection Algorithms

This section discusses the community detection algorithms used in this study, specifically *k-means clustering* and hierarchical clustering algorithms.

2.1.1 K-Means Clustering. K-Means clustering works as follows: given n nodes, k number of nodes are selected at random to serve as the initial centroids of k clusters. Each of the n nodes is then compared against all centroids and assigned to the cluster of the closest centroid to itself. New centroids are computed per cluster; each node is then reassigned to a cluster based on its distance from the new centroids. The algorithm ends when there is no change of clusters among all the n nodes. These clusters are the detected communities.

2.1.2 Hierarchical Clustering. Hierarchy-centric algorithms come in two forms: divisive and agglomerative. Divisive clustering places the entire set of nodes in one set; each set is then divided into two until each community only has one member. The division removes the node with the lowest edge betweenness since that node is most likely the node connecting two communities. Edge betweenness is defined as the number of shortest paths that pass along one edge. Agglomerative clustering starts with each node in their own community and communities are joined if they increase the overall modularity of the set of communities. The difference of the divisive and agglomerative clustering algorithms can be seen in Figure 2, wherein C1-C5 each refer to a community.

2.2 Similarity Parameters

[21] defines similarity measures for the *hashtag*, *following* and *retweeting* Twitter features.

Hashtag similarity is given by

$$sim_{hashtag}(i,j) = \sum_{k=1}^{n} \left(\left(1 - \left| \frac{N_{ik}}{|H_i|} - \frac{N_{jk}}{|H_j|} \right| \right) \frac{N_{ik} + N_{jk}}{|H_i| + |H_k|} \right) \quad (1)$$

where N_{ik} is the number of times user v_i used the hashtag k while H_i is the total hashtags used by v_i.

Following similarity is given by

$$sim_{follow}(i,j) = \frac{c_{friend}}{\sqrt{|Fr_i|\,|Fr_j|}} + \frac{c_{follower}}{\sqrt{|Fo_i|\,|Fo_j|}} \quad (2)$$

$|Fr_i|$ is the total number of users v_i follows. $|Fo_i|$ is the total number of users that follow v_i. c_{friend} represents the two users' common friends. $c_{follower}$ represents the two users' common followers.

Retweeting similarity is given by

$$sim_{retweet}(i,j) = \frac{c_{retweet}}{\sqrt{|R_i|\,|R_j|}} + \frac{n_{ij} + n_{ji}}{|R_i|\,|R_j|} \quad (3)$$

R_i is the number of users whom v_i retweet. $c_{retweet}$ is the number of users both v_i and v_j retweet. n_{ij} is the number of times v_i retweeted v_j and n_{ji} is the inverse case.

Mention similarity is given by

$$sim_{mention}(i,j) = \frac{c_{mention}}{\sqrt{|R_i|\,|R_j|}} + \frac{n_{ij} + n_{ji}}{|R_i|\,|R_j|} \quad (4)$$

R_i is the number of users whom v_i mention. $c_{mention}$ is the number of users both v_i and v_j mention. n_{ij} is the number of times v_i mention v_j and n_{ji} is the inverse case.

Another similarity measure is the cosine similarity. Cosine similarity is computed as the similarity of the two vectors. Given two vectors A with elements a_i, representing user A, and B with elements b_i, both vectors of length n, representing user B, cosine similarity is given as

$$cosine\ similarity = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2}} \quad (5)$$

Cosine similarity can be defined based on the features it will be used on.

2.3 Community Evaluation Metrics

This section discusses the metrics used to evaluate the communities, namely *modularity* and the *Davies-Bouldin Index (DBI)*.

2.3.1 Modularity. Modularity is used for measuring the strength of communities by comparing the strength of connections within a community to the strength of random connections between nodes [6]. This uses Equation 6.

Modularities approaching 1 indicate strong community structures; these values usually range from 0.3 to 0.7. However, a modularity of 0 would indicate communities that are only as good as randomly produced ones [6].

Modularity is given by

$$Q = \frac{1}{2m} \sum_{l=1}^{k} \sum_{i \in C_l, j \in C_l} (A_{ij} - \frac{d_i d_j}{2m}) \tag{6}$$

where m is the number of edges, d_i is the degree of node v_i, C_l being the lth community, and A_{ij} being the value in the adjacency matrix for node v_i and v_j [18] .

2.3.2 Davies-Bouldin Index. A common drawback to cluster algorithms is their dependence on user-set parameters which affect the algorithms' performance. The Davies-Bouldin Index (DBI) overcomes this difficulty as the user is only required to specify the distance and dispersion measure to be used [8]. This measure can be used to evaluate any clustering algorithm using the quantities and features in the dataset. The measure is symmetric and non-negative.

The DBI is computed as follows:

$$DB = \frac{1}{N} \sum_{i=1}^{N} D_i \qquad D_i = \max_{j \neq i} R_{i,j} \qquad R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \tag{7}$$

N is the number of clusters and D_i is the maximum of all $R_{i,j}$. $R_{i,j}$ is a measure of how good the clustering scheme is for cluster i and j where S_i and S_j are measures of scatter within cluster i and j, respectively. $M_{i,j}$ is a measure of separation between the aforementioned clusters.

The measure of scatter denoted as S_i can be computed in many ways. The following formula is an example of how the S_i can be computed:

$$S_i = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^q \right)^{\frac{1}{q}} \tag{8}$$

where T_i is the size of the cluster, A_i is the centroid of the cluster, X_j is an "n"-dimensional vector assigned to the cluster and q is a value that shall characterize the kind of distance of S_i will be between vectors and the centroid.

The measure of separation denoted as $M_{i,j}$ can also be computed in many ways. The following formula provides an instance on how $M_{i,j}$ can be computed:

$$M_{i,j} = \|A_i - Aj\|_p = \left(\sum_{k=1}^{n} |a_{k,i} - a_{k,j}|^p \right)^{\frac{1}{p}} \tag{9}$$

where n is the number of dimensions the centroids have, $a_{k,i}$ is the kth element of n-dimensional centroid A_i, $a_{k,j}$ is the kth element of n-dimensional centroid A_j, and p is a value that shall characterize the kind of distance of $M_{i,j}$ will be between centroids.

A lower value for the DBI means the "tightness" inside each cluster and the separation of the clusters are better which makes clustering better.

3 DESCRIPTION OF DATASET

Two datasets were crawled: one that represents users that talked about the Electronic Entertainment Expo (E3) conducted on June 13-15, 2017, and another that represents users from different Philippine universities. These datasets were selected since the authors had an expectation concerning the types of communities that would be formed by the algorithms.

3.1 E3 Dataset

Users were crawled based on companies they talked about. Since companies are most mentioned in posts, tweets were searched by keywords, with these keywords being the companies' names. While streaming tweets, the IDs of the users who posted these tweets were noted, and these users represented the company whose keyword they were retrieved with. Each company keyword was streamed until 250 users per company were obtained, for a total of 2500 users over 10 companies.

Some problems were encountered while building this dataset. First, retweets had to be filtered out because we wanted to prioritize users who published their own content, as opposed to simply echoing that of others. Second, tweets that mentioned @Youtube were removed, as Youtube has a feature that allows a user to automatically publish a tweet whenever he clicks the Like button on a Youtube video.

3.2 Philippine University Dataset

Users were crawled via keyword; via the Twitter Developer API, crawling users via keyword searched for users ala the"Find People" feature on Twitter. Given the keyword, Twitter searches for and returns users who have the keyword appear in their names or profile descriptions. Keywords that reflected five Philippine universities were used ("UP", "DLSU", "ADMU", "UST", and "Mapua"). A total of 1000 users was obtained, 200 from each university. An issue encountered with this dataset was that two keywords, namely "UP" and "UST", could not be used, as the users they returned were not reflective of these universities. The search with the keyword "UP" returned users with the word "up" in their Twitter profile descriptions. Likewise, searching with the "UST" keyword returned users that had "ust" in their names e.g. "Justin". The keywords had to be changed to "Diliman" and "Thomasian", respectively. This resolved the problem.

4 EXPERIMENT DESIGN

4.1 E3 Dataset Experiment

The first experiment used the E3 dataset mentioned in Section 3.1. Users were selected based on the companies they talked about. Ten companies were selected, namely Microsoft, Sony, Nintendo, Ubisoft, Bethesda, Naughty Dog, Atlus, EA, Kojima, and Rockstar. 250 users were obtained per company, for a total of 2500 users.

4.1.1 Results. Each combination of algorithm, feature, and similarity measure was run thrice, except combinations involving Agglomerative Hierarchical Clustering, as this is a deterministic algorithm, and only needs to be run once. Each time, the modularity, DBI, and number of communities generated were recorded. The results over the three runs of each combination can be seen in Tables 1 through

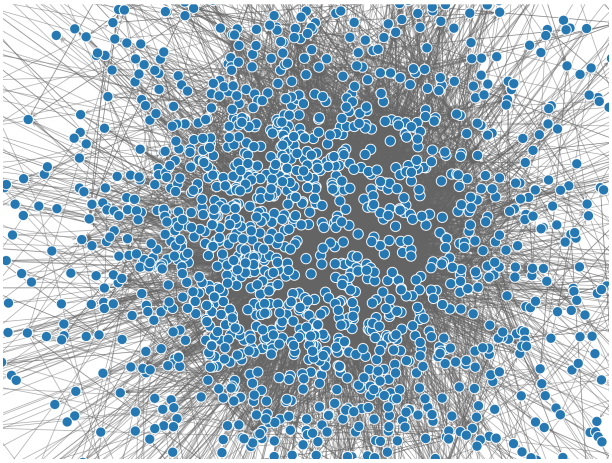


Figure 3: Sample graph structure for Dataset 1 (*Following* Feature)

2. The percent of users trimmed per feature is seen in Table 3. A visualization of the resulting community structure can be seen in Figure 3.

4.1.2 Analysis. Analyzing the performance across the different features, while all of them very rarely split the dataset into communities, *mentions* split the dataset the least, splitting only five times, and the *following* feature split the most, splitting 13 times. The *hashtags* feature split 10 times, and the *retweeting* feature split seven times. Analyzing the modularity scores, the percentage of users trimmed is directly proportional to the average modularity scores generated. The *retweeting* and *mentions* features generate higher average modularity scores in exchange for trimming a majority of the dataset. The *following* and *hashtags* features, on the other hand, do not obtain as high modularity scores but trim fewer users.

Generally speaking, the algorithms were rarely able to split the dataset into communities. 73% (93/128) of the runs were able to only generate one community. A possible explanation is due to a property of the dataset, which is when the average Twitter user goes to attend E3, or even just watches it online at home, he is not likely going to talk about just one company and forget the rest. Instead, he is much more likely to talk about most, if not all of the companies that presented at the conference. The resulting graph structure, as seen in Figure 3, supports this hypothesis, as, regardless of feature, it was too dense in the sense that there were too many edges between all the users for any of the Twitter features, for any of the algorithms to be able to perform any meaningful community detection on.

4.2 Philippine University Dataset Experiment

The second experiment used a dataset of 1000 Twitter users. The data was focused on five universities in Metro Manila, as mentioned in Section 3.2. Since the users gathered had a priori information about them, which was the university that they associated with, the algorithms should be able to detect these communities, and group the users based on this a priori information.

4.2.1 Results. Each combination of algorithm, feature, and similarity measure was run three times, except combinations involving *agglomerative hierarchical clustering*, as this is a deterministic algorithm, and only needs to be run once. Each time, the *modularity*, *DBI*, and *number of communities generated* were recorded. The results over the three runs of each combination can be seen in Tables 4 through 5. The percent of users trimmed per feature is seen in Table 6.

Each of the combinations generated communities of varying quality. [6] states that a modularity that is greater than 0.3 indicates a good community structure. The *retweeting* feature scored the highest modularity. However, this trimmed over 60% of the original dataset from the very beginning. The *mentions* feature came in second, but 47% of users had to be trimmed first. The *following* feature came in third, producing high modularity values without having to trim as many users (only 19% were trimmed). The *hashtags* feature performed the poorest of the four, causing more users to be trimmed down than for the *following* feature (29%), while also not reaching the 0.3 threshold specified by [6].

Looking at the community count, however, the *following* feature outperformed the rest of the features, as it was the only feature that was able to be consistently split into multiple communities. The *retweeting* feature came in second, splitting on two of the four algorithms. The *mentions* feature came in third, splitting on one of the four algorithms. The *hashtags* feature performed the poorest; it did not split on any of the four algorithms.

As mentioned in Section 2.3, a lower Davies-Bouldin Index (DBI) is indicative of a better community structure. For uniformity, all values referenced in this section are taken from the results obtained using the *cosine similarity* measure. For the *following* feature, the algorithm that performed the best in this regard was *divisive hierarchical clustering with simulated annealing*, producing a DBI of 1.399. For the *hashtags* feature, *k-means clustering with simulated annealing* performed the best, resulting in a DBI of 1.959. For the *retweeting* feature, *agglomerative hierarchical clustering with simulated annealing* performed the best, resulting in a DBI of 1.718. For the *mentions* feature, *divisive hierarchical clustering with simulated annealing* performed the best, yielding a DBI of 1.361. Based on the DBI, the algorithm that produced the most consistent communities was divisive hierarchical clustering with simulated annealing.

The Davies-Bouldin Index values that resulted from the combinations of algorithms and features are all above 1 but below 2, except for *divisive hierarchical clustering with simulated annealing* with *hashtags* feature using cosine similarity which yielded a DBI of 2.051. These values indicate some similarity between the communities, enough to combine them. This similarity could have resulted from the dataset that was used as most of the users were students from select universities in Metro Manila. It is because of this that it was likely that there were links between different students from different universities in the dataset.

For the *k-means* algorithm, the only feature that produced more than one community was the *following* feature. For *k-means* with Simulated Annealing, the *following* and *hashtags* features were able to split the communities. For standard similarity, the *mentions* feature was also able to split the communities during one of the three runs. For *divisive hierarchical clustering*, the *following* and *retweeting* feature were the only features that produced more than

Table 1: Results for Dataset 1 for Combinations of Each Algorithm using Similarity Measure according to [21]

		Following			*Hashtags*			*Retweeting*			*Mentions*		
		Modularity	DBI	# of communities	Modularity	DBI	# of communities	Modularity	DBI	# of communities	Modularity	DBI	# of communities
K-Means	Run 1	0.479	1.911	2	0.251	–	1	0.724	–	1	0.694	–	1
	Run 2	0.491	1.823	5	0.251	–	1	0.724	–	1	0.694	–	1
	Run 3	0.481	1.826	5	0.251	–	1	0.724	–	1	0.694	–	1
	Average	0.484	1.853	4.000	**0.251**	–	1.000	0.724	–	1.000	**0.694**	–	1.000
DHC	Run 1	0.493	1.822	5	0.251	–	1	0.746	1.329	5	0.694	–	1
	Run 2	0.494	1.832	4	0.251	–	1	0.724	–	1	0.694	–	1
	Run 3	0.501	1.819	5	0.251	–	1	0.724	–	1	0.694	–	1
	Average	0.496	1.824	4.667	**0.251**	–	1.000	**0.727**	1.329	2.333	**0.694**	–	1.000
AHC	Run 1	0.668	–	1	0.25	0.959	2	0.724	–	1	0.694	–	1
	Run 2	–	–	–	–	–	–	–	–	–	–	–	–
	Run 3	–	–	–	–	–	–	–	–	–	–	–	–
	Average	**0.668**	–	1.000	0.25	0.959	2.000	0.724	–	1.000	**0.694**	–	1.000
AHC SA	Run 1	0.043	1.228	461	0.251	–	1	0.059	1.581	640	0.694	–	1
	Run 2	0.034	1.209	563	0.251	–	1	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.251	–	1	0.59	1.846	69	0.069	1.596	590
	Average	0.248	1.219	341.667	**0.251**	–	1.000	0.458	1.7135	236.667	0.486	1.596	97.333
K-M SA	Run 1	0.668	–	1	0.251	–	1	0.724	–	1	0.568	1.99	2
	Run 2	0.539	1.943	3	0.251	–	1	0.724	–	1	0.568	1.99	2
	Run 3	0.619	1.912	2	0.251	–	1	0.556	1.989	2	0.694	–	1
	Average	0.609	**1.928**	2.000	**0.251**	–	1.000	0.668	**1.989**	1.333	0.61	**1.99**	1.667
DHC SA	Run 1	0.607	1.226	152	0.188	1.889	323	0.724	–	1	0.47	1.526	95
	Run 2	0.668	–	1	0.237	1.635	189	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.231	1.725	209	0.724	–	1	0.579	1.503	141
	Average	0.648	1.226	51.333	0.219	**1.750**	240.333	0.724	–	1.000	0.581	1.515	79.000

one community; among the two, the *retweeting* feature performed better. For *divisive hierarchical clustering with simulated annealing*, all combinations of Twitter features and similarity measures yielded more than one community except for *mentions* and *standard similarity*. For classical *agglomerative hierarchical clustering*, only the *following* feature returned more than one community. For *agglomerative hierarchical clustering with simulated annealing*, only the *hashtags* feature failed to produce more than one community. Among the remaining features, the *following* feature performed the best. It would then appear that out of the four algorithm's results, the feature that produced the tightest community structure was *following*. Also, generally, *standard similarity* resulted in lower DBI than *cosine similarity*. Combining the three exemplars, *DHC with simulated annealing* and the *following* feature on *standard similarity*, yielded a DBI of 1.205, which is the lowest DBI in the second dataset.

4.3 General Observations

The algorithms implemented can be split into two groups - the algorithms with and without simulated annealing (SA). The algorithms with SA tend to generate more communities than those without. The algorithms with SA tend to generate a small number of large communities and a large number of communities with only one user. These communities represented the users that did not have strong connections with anyone else in the dataset; the algorithms with SA were able to identify these users and place them in their own communities, while those without SA cluster them together with the other larger communities.

Incorporating different similarity measures, like the measures defined by [21] and the cosine similarity measure, there is no significant difference between the modularity and DBI scores of the generated communities, regardless of which similarity measure is used. In short, the communities generated were similar to each other. The observations regarding the effects the different features had on the communities formed still held regardless of what similarity measure was used, and among different features using the same similarity measure. This shows that the difference in community characteristics of *hashtags*, as opposed to *following*, *retweeting*, and *mentions*, is a result of the features themselves, and not of the similarity measure used.

Through the visualization module, qualitative analysis can be performed. The word clouds for the community tweets and profile descriptions represent what the users in each community frequently talk about, and how they describe themselves, respectively. By analyzing these word clouds, one can identify the characteristics of the users of each community and can validate/invalidate the a priori expectations that he may have regarding a certain dataset. For example, the a priori expectation that the dataset for the second experiment would cluster students based on the universities they represented can be validated by the tweet word clouds seen in Figure 4. This clustering is even further still reinforced by the profile word clouds, seen in Figure 5.

Table 2: Results for Dataset 1 for Combinations of Each Algorithm using Cosine Similarity

		Following			*Hashtags*			*Retweeting*			*Mentions*		
		Modularity	DBI	# of communities	Modularity	DBI	# of communities	Modularity	DBI	# of communities	Modularity	DBI	# of communities
K-Means	Run 1	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 2	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Average	**0.668**	–	1.000	**0.251**	–	1.000	**0.724**	–	1.000	**0.694**	–	1.000
DHC	Run 1	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 2	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Average	**0.668**	–	1.000	**0.251**	–	1.000	**0.724**	–	1.000	**0.694**	–	1.000
AHC	Run 1	0.668	–	1	0.25	0.959	2	0.724	–	1	0.694	–	1
	Run 2	–	–	–	–	–	–	–	–	–	–	–	–
	Run 3	–	–	–	–	–	–	–	–	–	–	–	–
	Average	**0.668**	–	1.000	0.25	0.959	2.000	**0.724**	–	1.000	**0.694**	–	1.000
AHC SA	Run 1	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 2	0.497	1.471	848	0.251	–	1	0.251	0.981	2	**0.694**	–	1
	Run 3	0.668	–	1	0.251	–	1	0.567	1.791	81	0.694	–	1
	Average	0.611	**1.471**	283.333	**0.251**	–	1.000	0.514	1.386	28.000	**0.694**	–	1.000
K-M SA	Run 1	0.668	–	1	0.19	1.917	4	0.724	–	1	0.694	–	1
	Run 2	0.668	–	1	0.215	1.919	2	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.251	–	1	0.476	1.966	3	0.694	–	1
	Average	**0.668**	–	1.000	0.218	**1.918**	2.333	0.641	**1.966**	1.667	**0.694**	–	1.000
DHC SA	Run 1	0.668	0.998	2	0.184	0.941	607	0.724	–	1	0.694	–	1
	Run 2	0.668	–	1	0.169	0.928	567	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.161	0.882	917	0.724	–	1	0.694	–	1
	Average	**0.668**	0.998	1.333	0.171	0.917	697.000	**0.724**	–	1.000	**0.694**	–	1.000

Table 3: Users Trimmed for Each Feature for Dataset 1

	% of Users Trimmed
Following	22.40%
Hashtags	13.80%
Retweeting	61.12%
Mentions	47.20%

Looking at the graph structure, both high-level among communities and low-level for a specific community, one can see how similar nodes are to each other, by looking at how close or how far away they are from one another. The visualization module can also identify outliers, as these would be the nodes that have longer edges between them and other nodes, signifying low similarity, or having few incoming/outgoing edges to the rest of the graph, signifying low connectivity to the rest of the dataset. For example, looking at the graph structures in Figure 6, one can easily see that the largest and most similar nodes tend to cluster near the center of the graph, and the further away a node is from the center, the less similar it is to the rest of the nodes, and the nodes that are nearest to the boundaries of the graph are most likely to be outliers.

Based on the results obtained, it was evident that using different features produced communities of different qualities. For example, clustering based on following generated communities that closely resembled the users' universities, while clustering based on hashtags generated communities based on topics that were common to users within them, regardless of which university these users came from. While the latter did not reflect the a priori expectations from the dataset, this did not necessarily mean that these communities were wrong or invalid.

Another insight that was made from the results is that modularity is not the only measure of the quality of the generated communities. For example, the modularity scores obtained by the *hashtags* feature were significantly lower than those of the rest of the features. However, this does not automatically mean that the communities they produced were any less valid, as they clustered the users based on their interests instead. The purpose of this study was to show that communities can be different, but still be equally valid.

5 CONCLUSION AND FUTURE WORK

This study explored the effect of various combinations of community detection algorithms and similarity parameters has on the communities formed by said algorithms. The algorithm typically

Table 4: Results for Dataset 2 for Combinations of Each Algorithm using Similarity Measure according to [21]

		Following			*Hashtags*			*Retweeting*			*Mentions*		
		Modularity	DBI	# of communities	Modularity	DBI	# of communities	Modularity	DBI	# of communities	Modularity	DBI	# of communities
K-Means	Run 1	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 2	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Average	**0.668**	–	1	**0.251**	–	1	**0.724**	–	1	**0.694**	–	1
DHC	Run 1	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 2	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Average	**0.668**	–	1	**0.251**	–	1	**0.724**	–	1	**0.694**	–	1
AHC	Run 1	0.668	–	1	0.25	0.959	2	0.724	–	1	0.694	–	1
	Run 2	–	–	–	–	–	–	–	–	–	–	–	–
	Run 3	–	–	–	–	–	–	–	–	–	–	–	–
	Average	**0.668**	–	1	0.25	0.959	2	**0.724**	–	1	**0.694**	–	1
AHC SA	Run 1	0.668	–	1	0.251	–	1	0.724	–	1	0.694	–	1
	Run 2	0.497	1.471	848	0.251	–	1	0.251	0.981	2	0.694	–	1
	Run 3	0.668	–	1	0.251	–	1	0.567	1.791	81	0.694	–	1
	Average	0.611	**1.471**	283.333	**0.251**	–	1	0.514	1.386	28	**0.694**	–	1
K-M SA	Run 1	0.668	–	1	0.19	1.917	4	0.724	–	1	0.694	–	1
	Run 2	0.668	–	1	0.215	1.919	2	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.251	–	1	0.476	1.966	3	0.694	–	1
	Average	**0.668**	–	1	0.218	**1.918**	2.333	0.641	**1.966**	1.667	**0.694**	–	1
DHC SA	Run 1	0.668	0.998	2	0.184	0.941	607	0.724	–	1	0.694	–	1
	Run 2	0.668	–	1	0.169	0.928	567	0.724	–	1	0.694	–	1
	Run 3	0.668	–	1	0.161	0.882	917	0.724	–	1	0.694	–	1
	Average	**0.668**	0.998	1.333	0.171	0.917	697	**0.724**	–	1	**0.694**	–	1

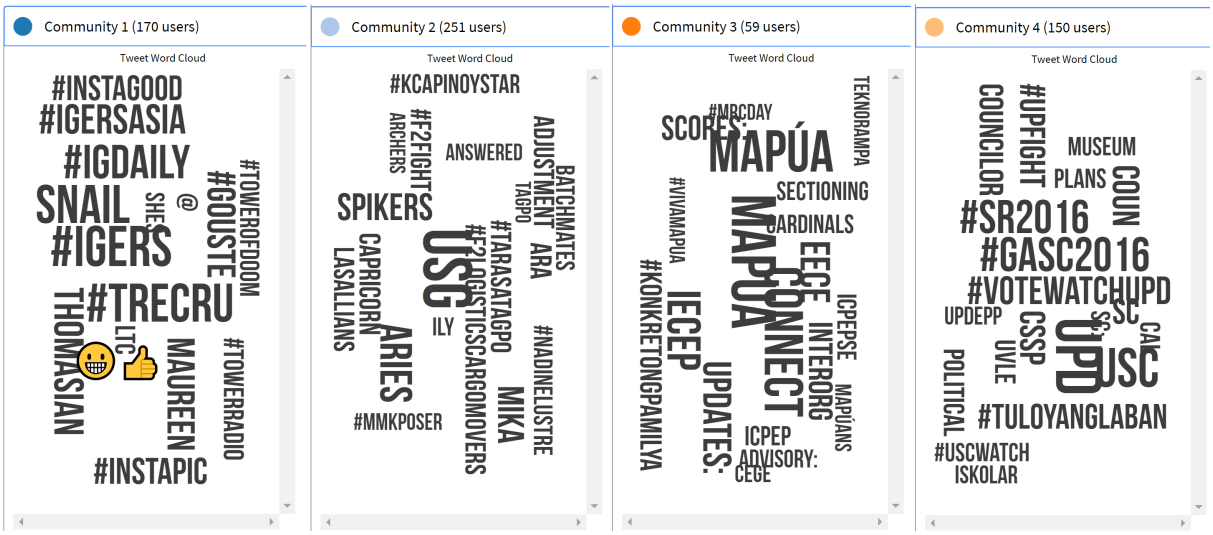


Figure 4: Sample Tweet Word Clouds for the communities generated by the Philippine universities dataset

dictated the overall structure of the communities; algorithms augmented with simulated annealing tend to generate a small number of large communities and several communities with only one user. Changing the similarity measures informed the types of communities formed; using the *following*, *mentions*, and *retweeting* features generally resulted in communities informed by explicit user interactions in the social network. Using the *hashtags* feature formed communities based on the topics or interests the users were discussing.

Table 5: Results for Dataset 2 for Combinations of Each Algorithm using Cosine Similarity

		Following			*Hashtags*			*Retweeting*			*Mentions*		
		Modularity	DBI	# of communities	Modularity	DBI	# of communities	Modularity	DBI	# of communities	Modularity	DBI	# of communities
K-Means	Run 1	0.459	1.949	2	0.251	–	1	0.717	–	1	0.584	–	1
	Run 2	0.468	1.93	3	0.251	–	1	0.717	–	1	0.584	–	1
	Run 3	0.45	1.938	3	0.251	–	1	0.717	–	1	0.584	–	1
	Average	0.459	**1.939**	2.667	**0.251**	–	1	0.717	–	1	0.584	–	1
DHC	Run 1	0.469	1.81	6	0.251	–	1	0.717	–	1	0.593	1.97	2
	Run 2	0.468	1.917	4	0.251	–	1	0.717	–	1	0.583	–	1
	Run 3	0.469	1.729	5	0.251	–	1	0.733	1.743	2	0.597	1.968	2
	Average	**0.469**	1.819	5	**0.251**	–	1	**0.722**	1.743	1.333	**0.591**	**1.969**	1.667
AHC	Run 1	0.45	1.409	80	0.251	–	1	0.717	–	1	0.584	–	1
	Run 2	–	–	–	–	–	–	–	–	–	–	–	–
	Run 3	–	–	–	–	–	–	–	–	–	–	–	–
	Average	0.45	1.409	80	**0.251**	–	1	0.717	–	1	0.584	–	1
AHC SA	Run 1	0.45	1.409	80	0.251	–	1	0.346	1.74	49	0.529	1.894	50
	Run 2	0.436	1.411	77	0.251	–	1	0.583	1.721	24	0.529	1.894	50
	Run 3	0.45	1.409	80	0.251	–	1	0.323	1.694	84	0.393	1.803	113
	Average	0.445	1.410	79	**0.251**	–	1	0.417	1.718	52.333	0.484	1.864	71
K-M SA	Run 1	0.467	1.932	4	0.184	1.968	4	0.717	–	1	0.584	–	1
	Run 2	0.461	1.951	2	0.251	–	1	0.717	–	1	0.584	–	1
	Run 3	0.453	1.914	5	0.148	1.95	8	0.717	–	1	0.584	–	1
	Average	0.460	1.932	3.667	0.194	1.959	4.333	0.717	–	1	0.584	–	1
DHC SA	Run 1	0.406	1.189	57	0.186	2.125	59	0.717	–	1	0.468	1.984	2
	Run 2	0.429	1.105	44	0.225	1.978	2	0.636	1.961	2	0.421	1.031	93
	Run 3	0.441	1.903	6	0.251	–	1	0.717	–	1	0.507	1.067	47
	Average	0.425	1.399	35.667	0.221	**2.052**	20.667	0.69	**1.961**	1.333	0.465	1.361	47.333

Table 6: Users Trimmed for every Feature for Dataset 2

	% of Users Trimmed
Following	19.50%
Hashtags	29.00%
Retweeting	65.50%
Mentions	47.20%

For future work, a standardized, streamlined method for obtaining data may be developed. Other algorithms such as *fast greedy optimization of modularity* [3] may be implemented, as well as other features such as direct messaging on Twitter or term frequency in tweets. The work may also be applied to social networks other than Twitter, which may use different features. Multiple features may also be used simultaneously when detecting communities to observe relationships between different combinations of features. Additionally, incorporating thresholds when measuring the similarity between users could be considered, to allow weaker connections to be removed, highlighting stronger connections. Moreover, allowing the system to detect features on its own based on the properties of the users in the data, may be explored.

In the visualization module, future work can present a more detailed graphical representation of the community structure of the network without compromising the performance of the user interface for larger datasets to better view the connections between communities rather than just within them, and provide more information for analysis of these communities after their generation. The visualization could also have a feature to select or deselect communities to view; this could be done to view specific communities' relationships with each other more clearly or to remove communities that may be deemed as outliers from the graph being viewed.

REFERENCES

[1] Benjamin R. C. Amor, Sabine I. Vuik, Ryan Callahan, Ara Darzi, Sophia N. Yaliraki, and Mauricio Barahona. [n.d.]. *Community detection and role identification in directed networks: Understanding the Twitter network of the care.data debate.* Chapter Chapter 5, 111–136. https://doi.org/10.1142/9781786340757_0005 arXiv:https://www.worldscientific.com/doi/pdf/10.1142/9781786340757$_0$005

[2] Bart Baesens. 2014. *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications* (1st ed.). Wiley Publishing.

[3] Mohamed Bakillah, Ren-Yu Li, and Steve H. L. Liang. 2015. Geo-located Community Detection in Twitter with Enhanced Fast-greedy Optimization of Modularity: The Case Study of Typhoon Haiyan. *Int. J. Geogr. Inf. Sci.* 29, 2 (Feb. 2015), 258–279. https://doi.org/10.1080/13658816.2014.964247

[4] John Bryden, Sebastian Funk, and Vincent AA Jansen. 2013. Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science* 2, 1 (2013), 3. https://doi.org/10.1140/epjds15

[5] Nan Cao, Lu Lu, Yu-Ru Lin, Fei Wang, and Zhen Wen. 2015. Socialhelix: visual analysis of sentiment divergence in social media. *Journal of visualization* 18, 2

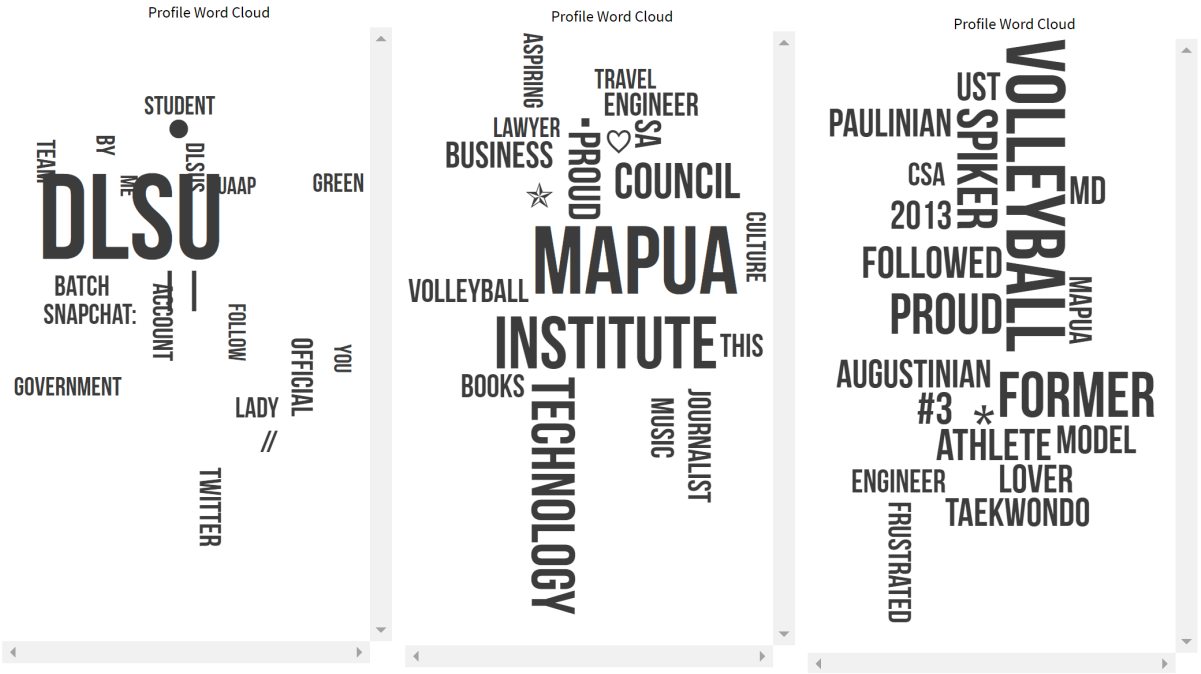


Figure 5: Sample Profile Word Clouds for the communities generated by the Philippine universities dataset

(2015), 221–235. https://doi.org/10.1007/s12650-014-0246-x

[6] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Phys. Rev. E* 70 (Dec 2004), 066111. Issue 6. https://doi.org/10.1103/PhysRevE.70.066111

[7] David Darmon, Elisa Omodei, and Joshua Garland. 2015. Followers are not enough: A multifaceted approach to community detection in online social networks. *PloS one* 10, 8 (2015). https://doi.org/10.1371/journal.pone.0134860

[8] D. L. Davies and D. W. Bouldin. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, 2 (April 1979), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[9] William Deitrick and Wei Hu. 2013. Mutually enhancing community detection and sentiment analysis on twitter networks. (2013). https://doi.org/10.4236/jdaip.2013.13004

[10] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (San Jose, California) *(WebKDD/SNA-KDD '07)*. Association for Computing Machinery, New York, NY, USA, 56–65. https://doi.org/10.1145/1348549.1348556

[11] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco, and Santo Fortunato. 2011. Finding statistically significant communities in networks. *PloS one* 6, 4 (2011).

[12] Kwan Hui Lim and Amitava Datta. 2012. Finding Twitter Communities with Common Interests Using Following Links of Celebrities. In *Proceedings of the 3rd International Workshop on Modeling Social Media* (Milwaukee, Wisconsin, USA) *(MSM '12)*. Association for Computing Machinery, New York, NY, USA, 25–32. https://doi.org/10.1145/2310057.2310064

[13] Kwan Hui Lim and Amitava Datta. 2012. *Following the Follower: Detecting Communities with Common Interests on Twitter*. Association for Computing Machinery, New York, NY, USA, 317–318. https://doi.org/10.1145/2309996.2310052

[14] Niall McCarthy. 2014. *Facebook versus twitter in numbers*.

[15] Thomas M. Mitchell. 1997. *Machine Learning* (1 ed.). McGraw-Hill, Inc., New York, NY, USA.

[16] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos. 2012. Community detection in social media. *Data Mining and Knowledge Discovery* 24, 3 (2012), 515–554. https://doi.org/10.1007/s10618-011-0224-z

[17] Warren Pearce, Kim Holmberg, Iina Hellsten, and Brigitte Nerlich. 2014. Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC Working Group 1 report. *PloS one* 9, 4 (2014). https://doi.org/10.1371/journal.pone.0094785

[18] Lei Tang and Huan Liu. 2010. *Community Detection and Mining in Social Media*. Vol. 2. 1–137 pages. https://doi.org/10.2200/S00298ED1V01Y201009DMK003 arXiv:https://doi.org/10.2200/S00298ED1V01Y201009DMK003

[19] Robert West, Hristo S. Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting Social Network Structure for Person-to-Person Sentiment Analysis. *Transactions of the Association for Computational Linguistics* 2 (2014), 297–310. https://doi.org/10.1162/tacl_a_00184 arXiv:https://doi.org/10.1162/tacl$_{a}$00184

[20] Jierui Xie. 2012. *Agent-based dynamics models for opinion spreading and community detection in large-scale social networks*. Ph.D. Dissertation. Rensselaer Polytechnic Institute.

[21] Yang Zhang, Yao Wu, and Qing Yang. 2012. Community discovery in twitter based on user interests. *Journal of Computational Information Systems* 8, 3 (2012), 991–1000. https://doi.org/10.1.1.465.9055

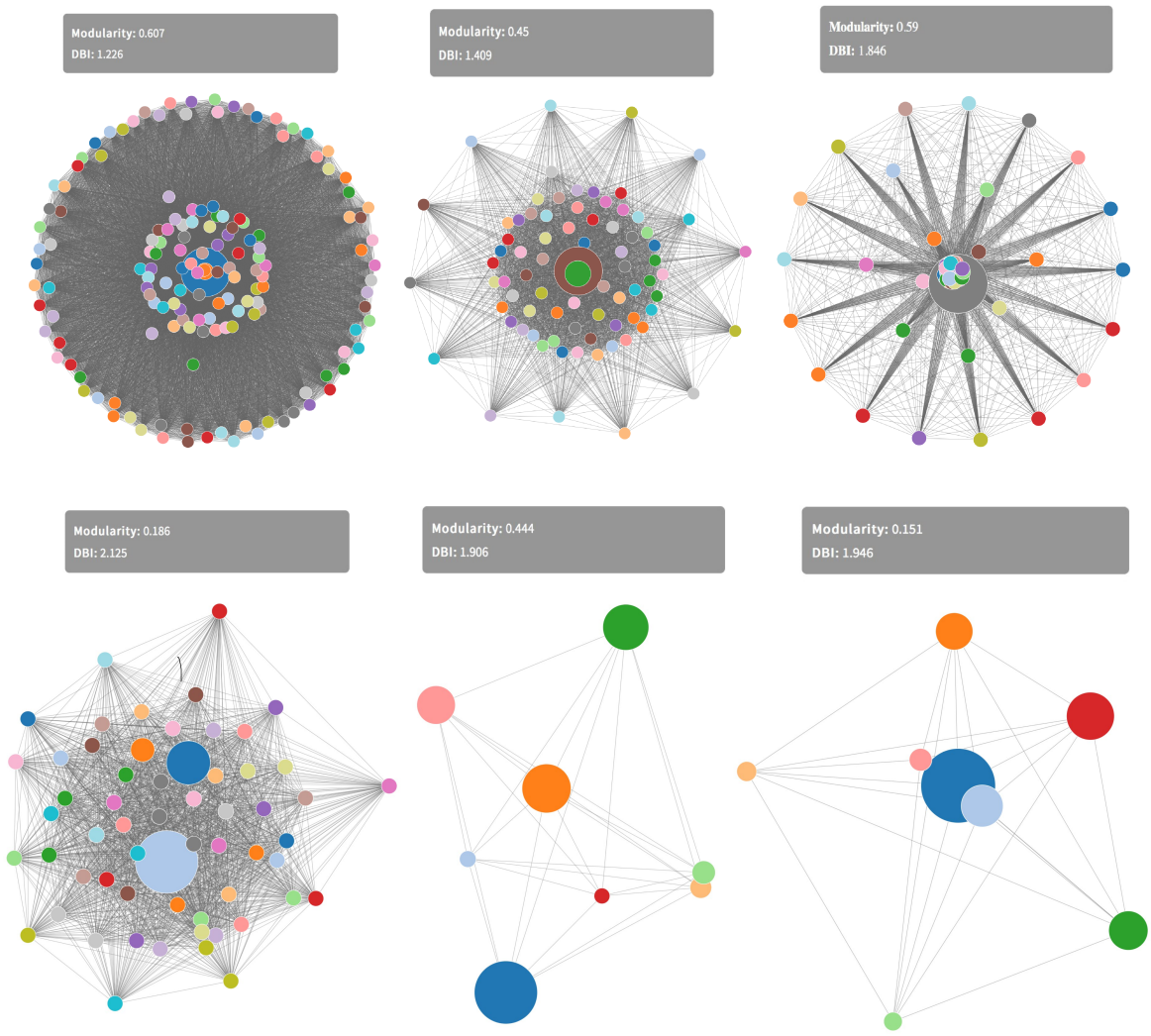


Figure 6: Sample graph structures for the communities generated by the Philippine universities dataset

Designing an Immersive VR Application Using Collective Memory for Dementia Therapy

Anne Lorelie M. Avelino[1], Paola Faith T. Simon[1], Paul Matthew L. Sason[1]
Richelle Ann B. Juayong[1], Jasmine A. Malinao[2], Veeda Michelle M. Anlacan[3]
Michael L. Tee[3], Gregg S. Lloren[4], Jaime D.L. Caro[1]

[1]Service Science and Software Engineering Lab
Department of Computer Science
University of the Philippines Diliman
Quezon City, Metro Manila, PH

[2]Headstart Business Solutions, Inc.
Quezon City, Metro Manila, Philippines

[3]College of Medicine
University of the Philippines Manila
Metro Manila, Philippines

[4] College of Communication, Arts and Design
University of the Philippines Cebu
Cebu City, Philippines
{amavelino,ptsimon,plsason,rbjuayong,vmanlacan,mltee,gslloren,jdlcaro}@up.edu.ph
jas_malinao@yahoo.com

ABSTRACT

There is currently a lack of research, facilities, services, and workforce that specifically cater to dementia in the Philippines. We are currently limited to drug medication and home care, with little development towards other potential ways of treatment and rehabilitation. The possible use of VR technology for dementia therapy is also not yet explored in the Philippine context. In this paper, we present an application design for an immersive virtual reality software. This software is intended as a supplementary therapy to help elderly patients with dementia. To incorporate a form of reminiscence therapy, we use the concept of collective memory as a scheme for personalization and an effective concept to help patients with dementia.

KEYWORDS

dementia, immersive virtual reality, collective memory

ACM Reference Format:
Anne Lorelie M. Avelino[1], Paola Faith T. Simon[1], Paul Matthew L. Sason[1], Richelle Ann B. Juayong[1], Jasmine A. Malinao[2], Veeda Michelle M. Anlacan[3], and Michael L. Tee[3], Gregg S. Lloren[4], Jaime D.L. Caro[1]. 2020. Designing an Immersive VR Application Using Collective Memory for Dementia Therapy . In *Proceedings of Workshop on Computation: Theory and Practice (WCTP2020).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

WCTP2020, November 21, 2020, Online Conference

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Dementia is an overall term for a collection of symptoms of cognitive decline that is severe enough to interfere with a person's daily activities [1]. This cognitive decline may refer to the loss of memory, trouble with understanding or using words, and difficulty in moving despite having adequate motor functions, and failure to identify or recognize objects. [1] Dementia comes in many forms, with Alzheimer's disease being the most common, accounting for 60-80% of cases. [2] [3] The memory and cognitive decline also often manifests as behavioral and mood changes, as well as the inability to perform daily living activities [3].

According to the World Alzheimer Report 2015 [4], the prevalence of Dementia in Southeast Asia was projected to increase by 236% from 2015 to 2050. The prevalence of Dementia especially in the Philippines was found to be high as of 2018 [5]. New cases of dementia are expected to grow exponentially. It is projected that more than a million dementia sufferers among senior citizens (age 60 years and older) in the Philippines are expected in 2040, which is nearly five times those in 2010. The number of new cases of dementia in 2015 is expected to more than triple by 2045 [6]. The World Health Organization (WHO) and the Alzheimer's Disease International (ADI) considered the disease a public health priority

[7], however, this is not the case in the Philippines. Dementia diagnosis and management in non-acute settings is not recognized as a major public health concern and older individuals with dementia are usually not prioritized in the face of limited family resources. Research and awareness is also generally lacking in the Philippines, with limited social and health care reform in dementia care [5].

Most types of dementia have treatments that could stop or reverse its progression. These treatments aim to alleviate the symptoms of dementia as well as raise the quality of life of the patient with dementia [2]. Treatments includes Pharmacological agents which are the drug medicines and supplements prescribed by the doctor. While the non-pharmacological interventions include exercise (such as aerobics and strength training), cognitive training, aromatherapy, music therapy, and massage [3] [8]. In the Philippines, drug medication is the most prominent treatment given to dementia patients [9]. Psychiatric therapies are offered but only by private institutions, and they are usually very costly as they come in multiple session package. Most dementia patients are taken care of by their families in their homes, but there are also home for the aged (HFTA) centers which can take care of the elderly, including those with dementia [9].

Another form of therapy for dementia patients is the so-called Reminiscence therapy which involves recollection and discussion of past events and experiences with a person or a group of people [10]. It has been in use as a therapy, particularly for older persons, aiming to increase general happiness and maintaining self-esteem [11]. This usually involves individual or group discussions about past memories facilitated by a therapist as performed in various studies [10], but can be more varied and tap into other senses like sound and smell [12]. As these studies show, the overall goal is to stimulate a strong reminiscence, preferably on the more positive memories.

Reminiscence therapy also involves the idea of *collective memory*, individual memories shared by a group of people. According to [13], society itself plays a role on how these memories are shaped. There can be different social cues or artifacts that can influence which memories are more remembered by groups of people, and it's dependent on what kind of environment or culture the people are in. In line with this, research also suggests that for older people, memories from their youth and early-adulthood (around ages 10-30) are the ones more remembered - a phenomenon called *reminiscence bump*. Recall of memories from this period is shown to be better when people were given word cues for reminiscence. [14] [15]

The usage of Virtual Reality (VR) Systems is emerging as one of the potential tools that can be used for dementia assessment and treatment, as it offers the possibility of performing activities or tasks in a Virtual Environment (VE) that can be modified to adapt to the needs of the patient [16]. It already has applications in the field of psychotherapy, such as in the treatment of phobias, and various medical uses such as post-stroke intervention [16]. Compared to traditional therapy, VR offers the benefit of reusable equipment, as the equipment used isn't limited to dementia therapy, but can also be used for other types of therapy [17]. It presents therapy in a more engaging and motivating manner for the patient, as compared to traditional therapy that may seem boring or repetitive [17]. It can allow remote treatment, which is especially important in rural areas, and it can provide increased privacy for the patient, as it can be used in an enclosed room [17].

There are already existing usage of Virtual Reality (VR) Technology in Alzheimer's disease (AD) applications (a review of these applications is given in [16]). Also, there are already existing VR applications for Dementia ([18–21]). However, in the Philippine context, the possible use of VR technology for dementia therapy is not yet explored. According to [22], there is currently a lack of research, facilities, services, and workforce that specifically cater to dementia in the Philippines. We are currently limited to drug medication and home care, with little development towards other potential ways of treatment and rehabilitation [9]. Moreover, we have not found existing literature that uses VR applications for dementia, locally and internationally, that considers integrating collective memory to incorporate reminiscence therapy in their applications. Since such therapy is a usual approach to helping dementia patients, incorporating collective memory may increase the effectiveness of such VR applications.

In this paper, we propose a design for an immersive VR software that will serve as a supplementary therapy for elderly patients with dementia. Its purpose is to improve their quality of life by slowing down the decline of their memory while enjoying at the same time. For our VR application, we use the idea of reminiscence bump. That is, we incorporate a reminiscence therapy shaped around the collective memory of a specific time period, with the idea that these memories would be the ones most remembered by our target demographic.

The outline of this paper proceeds as follows: In Section 2, we present our objectives and consideration when designing the VR application. In Section 3, we present our design, initial visualizations and test plans. We give our conclusions in Section 4.

2 OBJECTIVES AND CONSIDERATIONS

Our goal is to eventually develop an immersive VR software as a supplementary therapy that primarily aims to help elderly patients with dementia. The main goals we aim to implement in our application are: (a) To use collective memory as a personalization scheme for our application, (b) To provide memory-stimulating activities targeted to dementia patients to slow down their memory decline, and (c) To improve overall quality of life by providing a relaxing and enjoyable VR application.

To accomplish this, we consider a design that includes the following features:

(1) Personalized Collective Memory
A distinct feature in our application is the use of the collective memory of elderly people within the elements of the virtual environment that serves as a reminiscence therapy for the patients. This is for the experience to be more personal and relatable with the use of familiar elements from their past. Following the concept of reminiscience bump, elderly people tend to access more personal memories in their life from approximately 10-30 of age. With the basis that the vast majority of people with Alzheimer's dementia is age 65 or older [23], we have chosen to use popular culture in the Philippines during the 1960's to the 1980's. We chose to include elements from Philippine pop culture, specifically

targeting popular music, tv shows, movies/films, celebrities, sports and furniture.

(2) Patient Profile with Image Upload
We want to develop an application that can be used and reused by different users. As such, we shall maintain a patient profile for users of the application. A profile may contain a patient's personal information such as name, age, and patient ID. Profiling also allows tracking of the patient's scores in the activities for an overview of their performance (e.g. game scores). This will help the therapist analyze the progress of the patient's cognition.
Each profile can be associated with uploaded images to be used within the virtual environment. This will further support reminiscence therapy and also provide a deeper level of personalization. It is assumed that the selection and upload of images will be accomplished or guided by the therapist so that only the images that can trigger positive memories will be loaded in the application.

(3) Relaxing Scenery and Music
As with VR applications for dementia (.e.g [24]), we want to incorporate sensory therapy. In this way, dementia patients can engage their senses in visually and aurally relaxing sceneries that offer the possibility of alleviating anxiety and experience fascination.

(4) Cognitive Stimulating Activities
Difficulties with memory, language, problem-solving, and other cognitive skills are the general characteristics of a person with dementia. The decline in their memory worsens their condition that disrupts their daily life. [1]. In order to slow down the decline of memory, cognitive training or brain training is needed [20]. We shall include memory games and puzzles in the application that is proven to be beneficial to brain health and cognition among adults with dementia, especially Alzheimer's disease. [25]

Given below are key design concepts listed by [19] and exclusion criteria from the study [18] that should also be considered when developing a VR application intended for older users:

(1) *Visual and Aural Constraint* - possible VR-induced sickness and side-effects include motion-sickness or dizziness. To reduce or avoid these, visual effects must be constrained as not too bright, too flashy, and disorienting. Shaky and unstable camera must also be avoided. Sounds and music must not be too loud or noisy as it may be startling, confusing, and even scary for the elderly.

(2) *Realistic High-Quality Environment* - older people prefer a deeper level of realism to provide authenticity in their experience and to further immerse themselves in the virtual world. This means design models of each object in the virtual world would be as close as possible to reality, and resolution quality must be at least 2K (20148x1080).

(3) *Achievement-oriented* - older people also want goal-oriented tasks rather than open-ended ones as this reduces the overall complexity of the virtual game and help them to be more focused. Completion of these tasks also gives them a sense of achievement.

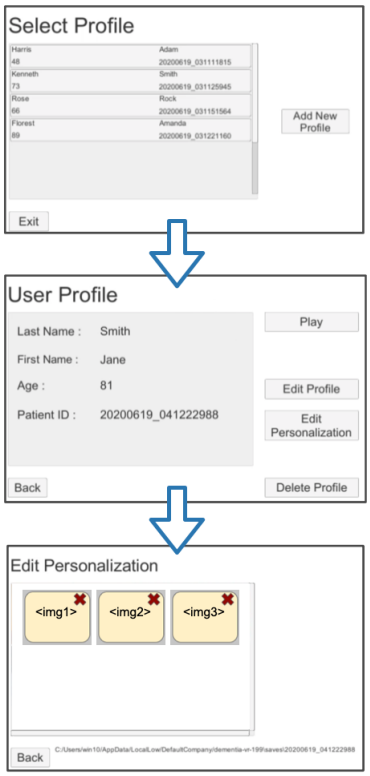


Figure 1: An example flow illustrating update of a patient profile.

(4) *User Interface Simplicity* - UI must be as simple as possible, with minimal as possible menus and buttons to reduce visual complexity.
Additionally, we have added one more constraint that is not much specified in most literature, but we think is an important thing to consider when using VR technology for the elderly:

(5) *Time Limit per Session* - the recommended usage of the application must be limited to a certain time, and/or add short breaks in between to minimize possible side-effects.

3 APPLICATION DESIGN

When the application has loaded, the user will be shown a menu screen where they can select, add, edit, and delete profiles. Each profile stores the following: (a) personal details of the user (e.g. name and age), (b) a generated number ID, (c) personal images they want to use within the game, and (d) their saved progress and settings. A user can upload their own images for a more personalized experience. It may be photos of their past and/or photos related to them personally, i.e. family member, pet, an item with sentimental value. These images will be displayed within the virtual environment and used for the mini-games. An example flow in the application illustrating updating of a profile is given in Figure 1. Once the profile is setup and selected, user can now enter the virtual environment.

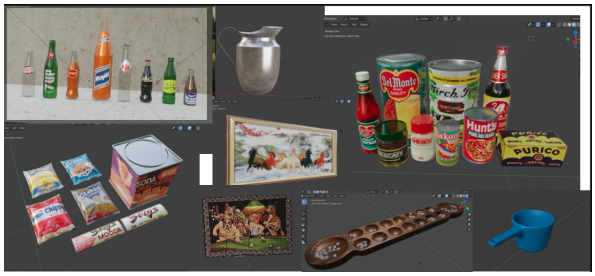


Figure 2: Food and items during the 1960s-1980s modeled in the VR environment

Figure 3: Top view of the virtual environment. The sala, leisure area, dining area, and receiving area is located in the lower right, upper right, upper left and lower left of the house, respectively. Outside the house is the garden backyard.

The virtual environment is mainly composed of a home with a garden backyard. The house is designed based on an average middle-class Filipino household in the 60s-80s (see Figure 3 for a top view of virtual environment design). The house consists of a sala, a dining area, a leisure area and a receiving area. Familiar pieces of furniture, e.g. common house paintings, tables with picture frames, dining table with familiar food/snacks, is also included in the home area. Given in Figure 2 are some example food and items in the 1960s-1980s that are found in the virtual home.

The game will be played through a first-person point of view for an immersive gameplay experience. The user can look and walk around the environment using the set controls. In the virtual environment, the user can navigate and interact with some objects. When an object is triggered, the object may be looked at (zoom in), held or released. This will enable the user to perform usual household routines. For example, the TV in the sala area can be switched on/off or change channel. Upon switching on, famous tv shows, commercials or news from 60s-80s will be played. The user can also interact with the radio in the leisure area, e.g. turn the radio on/off or change radio station . Upon switching it on, famous music from the mentioned year range will be played.

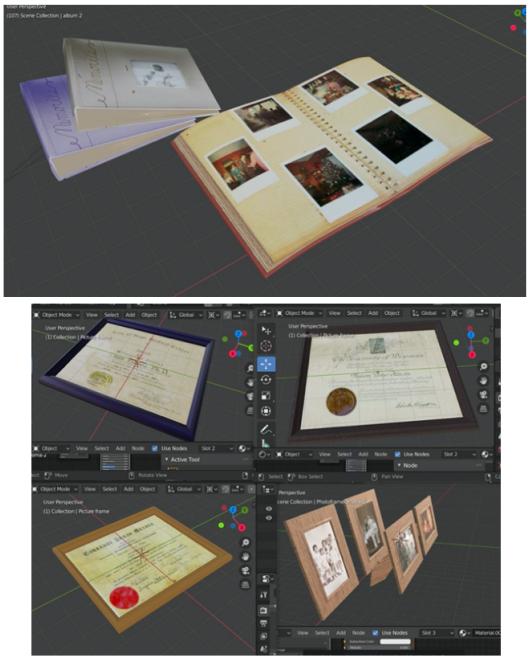


Figure 4: An illustration of how uploaded images will be displayed in the VR environment as picture frames and photo albums

Other interactive objects are designed so that the user can perform usual household chores. For example, the following activities can be performed by the user:

- In the garden area, the user can water the plants.
- In the sala, the user can switch the radio/tv on/off and change channels and turn the lamp on/off.
- In the dining area, the user can open the refrigerator and set plates on the table.
- In the receiving area, the user can dial the telephone.
- In any part of the house, the user can open/close doors and windows.

When a user uploads their own images (of their family members, pets, etc.), it will be displayed around the virtual environment in the form of picture frames and photo albums (see Figure 4). Designs of the sala, dining area, leisure area, receiving area and garden of the virtual environment are given in Figures 5, 6, 7, 8 and 9, respectively.

To incorporate cognitive stimulating activities, some objects in the virtual home trigger minigames. We are considering implementation of the ff minigames: (a) card matching, (b) light pattern memory game, (c) painting pieces puzzle and (d) family tree matching game. The cards in the receiving area, one channel in the TV, the paintings scattered around the house and the photo album in the leisure area are objects that trigger these minigames, respectively.

- Card Matching Game
  Card matching is a simple game that aims to exercise short-term visual memory (see: [26], [9], [27]). In this game, a list of cards are initially revealed to the user. Afterwards, the cards

Figure 5: The sala of the virtual environment. The painting and the TV are both interactable.

Figure 6: The dining area of the virtual environment. The cabinet is interactable.

Figure 7: The leisure area of the virtual environment. The photo albums and radio are interactable.

Figure 8: The receiving area of the virtual environment. The cards in the table trigger a minigame.

Figure 9: The backyard of the virtual environment. The pail and plants are interactable so that the household chore of watering plants can be accomplished by the user.

will be turned over and the user will have to select which pairs of cards are the same. A range of images can be used for the card matching game. For example, we can use generic pictures of famous snacks, celebrities or furnitures. We can also use pictures of popular or familiar objects present in the 60s-80s. Finally, we can select images from the ones uploaded in the user's profile. The choice of images can also be customized by the therapist.

- Light Pattern Memory Game
  In this game, there are four buttons of different colors that will be used. These buttons lights up. Thus, a pattern is formed by a sequence of lit up buttons. The user must therefore replicate the pattern by pressing down the buttons in order after the pattern plays. This game exercises short term visual memory, concentration and pattern matching. The level of difficulty increases as we increase the length of the light pattern.
- Painting Pieces Puzzle
  We shall adapt a game used in one VR application for dementia that is illustrated in Figure 10. In this game, painting puzzles are scattered around the house for the user to collect and piece together. Puzzles are good activities to exercise visual patterns.
- Family Tree Matching Game
  This game will only be possible provided images of family members are uploaded. This game is considered for the patient to exercise long term memory, specifically his/her autobiographical memory and reminiscence [15]), [21]. One of the pages of a photo album in the virtual home triggers this particular game. A family tree is displayed and the user will

Figure 10: An illustration of the painting pieces puzzle from [21]. The user finds pieces of puzzle around the house and piece them together.

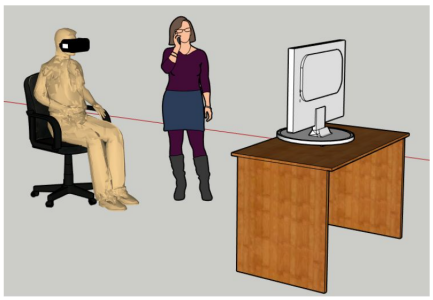


Figure 11: Proposed setup when using the application

have to put images of family members in their appropriate positions in the family tree.

The user profile will also record the scores of the user in these mini-games. This can be used by the therapist in evaluating the cognitive abilities of the user for testing purposes and general examination of the patient's cognition.

To incorporate sensory therapy, relaxing and calming background music and sounds are used throughout the game while the user navigate the virtual environment. The garden beside the house also serves as a relaxing scenery that the user can wander and explore. In developing the elements of the environment, we consider the overall theme to be nostalgic, with relaxing sceneries, many flowers and soft filters.

3.1 Setup and Implementation Specifics

Figure 11 provides an illustration of the proposed setup when using the application. The application is intended to be used with a head-mounted display (HMD). The user wears a head-mounted gear and sits on a chair while the therapist gains a preview of the VR environment seen by the user.

An initial demo of the application being developed and following the design mentioned above is given at https://sites.google.com/up.edu.ph/upd-dcs-s3lab/teachings/cs199_ay2019-2020. The application is developed using Unity for the game engine and Blender 3D for object modeling and creation.

There are several equipment being considered as of the moment, taking into account the cost and ease-of-use specifically for older users.

- Generic VR headset + handheld controller
  This option offers flexibility as the headset and controller can be bought according to a budget, if there are any monetary constraints. The controller in this option needs to have a joystick and at least one button that can be assigned as the "action button" - basically, the button that the user will be pressing in order to interact with the VR environment. Smartphone-based VR such as this is subject to the limitations of the phone-hardware. The phone used alongside the headset needs to have the minimum specifications needed to run a VR application: the OS has to be Android 4.1 or above; and the phone should have a gyro sensor. This minimum set of specifications may also change as the VR application is developed.
- Oculus Go
  This option is a VR Headset that doesn't require a phone to be used/inserted, but a phone is still needed for the download of the Oculus App. Wireless internet access is also required. Oculus Go comes with a single controller that can be used in either hand. This controller has a trackpad and a trigger button. One difficulty in the use of this equipment is mapping the movement controls - right now, we are looking at using the "scrolling" gesture on the trackpad to make the player move inside the environment.
- Oculus Rift S
  Similar to Oculus Go, this equipment doesn't need a phone to be user/inserted into the device. However, it requires a PC that follows some specific requirements [28]. Oculus Rift S comes with 2 controllers, one for each hand. Each controller has thumbstick, a trigger button and two buttons (X and Y buttons on the left-hand controller, A and B buttons on the right-hand controller). This VR headset is wired, which could possibly bother the user, or even hinder their movement. The controllers might be complicated to elders with dementia, as there are more things they can fiddle with on the controller. Also, this equipment is more costly as compared to the other alternative

We shall also explore the use of a Leap Motion Sensor to have a hands-free setup. The Leap Motion Sensor is mounted on a VR headset in order to track the player's hand movements. A research direction to consider in this case is how to create a hand gesture that feels intuitive for the user when they move inside the VR environment.

3.2 Testing Plan

The main feature of the application is the integration of collective memory. To test the effectiveness and impact of such feature, we will follow existing testing methods and metrics from [18] and [21] when the application is tested on dementia patients. However, before testing on actual patients, we will first conduct tests involving a group of healthy people (aged 18-70) with a background on the experiences and needs of a person with dementia (e.g. a relative or caregiver of someone with dementia, research students, or a

specialist that deals with dementia patients). For the test, we have two main goals: (a) to measure the software product quality by evaluating its functional suitability and usability following ISO/IEC 25010[29] (b) to test usability and user satisfaction of the VR application by collecting feedback from the users/testers based on their experience during and after using the application. Functional suitability measures the systems functional completeness, correctness, and appropriateness based on the requirements specifications while usability measures the system can be used by specified users to achieve specific goals with effectiveness, efficiency, and satisfaction.

4 CONCLUSION

In this paper, we discuss our initial design and test plans on developing a VR application that functions as a supplementary therapy for people with dementia. The design adapts features of exist-ing VR application, thus incorporating multiple therapy schemes, such as reminiscence therapy, sensory therapy, and game ther-apy. Additionally, this study provides a preliminary work on the use of collective memory in a VR application as part of remi-niscence therapy. An initial demo of the application being de-veloped is given at https://sites.google.com/up.edu.ph/ upd-dcs-s3lab/ teachings/cs199_ay2019-2020.

REFERENCES

[1] D. P. Chapman, S. M. Williams, T. W. Strine, R. F. Anda, and M. J. Moore. Dementia and its implications for public health. *Prev Chronic Dis*, 3(2):A34, Apr 2006.

[2] What is dementia? https://www.alz.org/alzheimers-dementia/what-is-dementia. Accessed: 2019-09-18.

[3] K. R. Scott and A. M. Barrett. Dementia syndromes: evaluation and treatment. *Expert Rev Neurother*, 7(4):407–422, Apr 2007.

[4] Dr. Maëlenn Guerchet Gemma-Claire Ali Dr. Yu-Tzu Wu Dr. Matthew Prina Martin Prince, Anders Wimo and Alzheimer's Disease International. World alzheimer report 2015: The global impact of dementia. Technical report, 2015.

[5] Jacqueline Dominguez, Ma. Fe de Guzman, Macario Reandelar, and Thien Kieu Thi Phung. Prevalence of dementia and associated risk factors: A population-based study in the philippines. *Journal of Alzheimer's Disease*, 63(3):1065–1073, May 2018.

[6] N. Ogena. Population aging and recent projections of dementia and alzheimer's disease in the philippines. In *18th Asia-Pacific Regional ADI Conference, November 18-21, 2015, Philippines.*, 2015.

[7] World Health Organization and Alzheimer's Disease International. Dementia: a public health priority. Technical report, 2012.

[8] F. Mulla, Eya Eya, E. Ibrahim, A. Alhaddad, R. Qahwaji, and R. Abd-Alhameed. Neurological assessment of music therapy on the brain using emotiv epoc. In *2017 Internet Technologies and Applications (ITA)*. IEEE, September 2017.

[9] MD Michelle M. Anlacan. Personal Communication. October 28, 2019.

[10] Maria Cotelli, Rosa Manenti, and Orazio Zanetti. Reminiscence therapy in dementia: A review. *Maturitas*, 72(3):203–205, July 2012.

[11] K. S. Latha, P. V. Bhandary, S. Tejaswini, and M. Sahana. Reminiscence therapy : An overview. *Middle East Journal of Age and Ageing*, 11(1):18–22, January 2014.

[12] Annie M.H. Chin. Clinical effects of reminiscence therapy in older adults: A meta-analysis of controlled trials. *Hong Kong Journal of Occupational Therapy*, 17(1):10–22, January 2007.

[13] Alin Coman, Adam D. Brown, Jonathan Koppel, and William Hirst. Collective memory from a psychological perspective. *International Journal of Politics, Culture, and Society*, May 2009.

[14] Khadeeja Munawar, Sara K. Kuhn, and Shamsul Haque. Understanding the reminiscence bump: A systematic review. *PLOS ONE*, 13:1–36, 12 2018.

[15] Jonathan Koppel and David C. Rubin. Recent advances in understanding the reminiscence bump: The importance of cues in guiding recall from autobiographical memory. *Current Directions in Psychological Science*, 25(2):135–140, April 2016.

[16] Rebeca I. García-Betances, María Teresa Arredondo Waldmeyer, Giuseppe Fico, and María Fernanda Cabrera-Umpiérrez. A succinct overview of virtual reality technology use in alzheimer's disease. *Frontiers in Aging Neuroscience*, 7:80, 2015.

[17] G. C. Burdea. Virtual rehabilitation – benefits and challenges. *Methods of Information in Medicine*, 42(05):519–523, 2003.

[18] Luma Tabbaa, Chee Siang Ang, Vienna Rose, Panote Siriaraya, Inga Stewart, Keith Jenkins, and Maria Matsangidou. Bring the outside in: Providing accessible experiences through vr for people with dementia in locked psychiatric hospitals. In *ACM CHI Conference on Human Factors in Computing Systems 2019*. ACM, May 2019.

[19] Panote Siriaraya and Chee Siang Ang. Developing virtual environments for older users: Case studies of virtual environments iteratively developed for older users and people with dementia. In *2017 2nd International Conference on Information Technology (INCIT)*. IEEE, November 2017.

[20] M. Intraraprasit, P. Phanpanya, and C. Jinjakam. Cognitive training using immersive virtual reality. In *2017 10th Biomedical Engineering International Conference (BMEiCON)*, pages 1–5, Aug 2017.

[21] G. Caggianese, A. Chirico, G. De Pietro, L. Gallo, A. Giordano, M. Predazzi, and P. Neroni. Towards a virtual reality cognitive training system for mild cognitive impairment and alzheimer's disease patients. In *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 663–667, May 2018.

[22] Shelley F. Dela Vega, Cynthia P. Cordero, Leah A. Palapar, Angely P. Garcia, and Josephine D. Agapito. Mixed-methods research revealed the need for dementia services and human resource master plan in an aging philippines. *Journal of Clinical Epidemiology*, 102:115 – 122, 2018.

[23] 2018 alzheimer's disease facts and figures.

[24] How a virtual reality forest helps alzheimer's patients. https://www.alzheimers.net/how-a-virtual-reality-forest-helps-alzheimers-patients/. Accessed: 2019-09-18.

[25] S. A. Schultz, J. Larson, J. Oh, R. Koscik, M. N. Dowling, C. L. Gallagher, C. M. Carlsson, H. A. Rowley, B. B. Bendlin, S. Asthana, B. P. Hermann, S. C. Johnson, M. Sager, A. LaRue, and O. C. Okonkwo. Participation in cognitively-stimulating activities is associated with brain structure and cognitive function in preclinical Alzheimer's disease. *Brain Imaging Behav*, 9(4):729–736, Dec 2015.

[26] G. Burdea, B. Rabin, D. Rethage, F. Damiani, J. S. Hundal, and C. Fitzpatrick. BrightArm^{TM} therapy for patients with advanced dementia: A feasibility study. In *2013 International Conference on Virtual Rehabilitation (ICVR)*. IEEE, August 2013.

[27] Tiffany Tong, Jonathan H. Chan, and Mark Chignell. Serious games for dementia. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press, 2017.

[28] Oculus Rift and Rift S Minimum Requirements and System Specifications. https://support.oculus.com/248749509016567/?locale=en_GB&ref=oculus-pc-app%23specs. Accessed: 2020-06-15.

[29] ISO/EIC 25010. https://iso25000.com/index.php/en/iso-25000-standards/iso-25010. Accessed: 2019-09-18.

Building a Corpus of Emotional Facial Expressions Towards the Development of an Affective Filipino Embodied Agent

Joshua Terence D. Del Barrio, Shan Lee C. Kim, Miguel Fernando C. Rivera, Judith Azcarraga*

College of Computer Studies

De La Salle University, Manila, Philippines

joshua_delbarrio@dlsu.edu.ph, shan_kim@dlsu.edu.ph, miguel_fernando_rivera@dlsu.edu.ph, judith.azcarraga@dlsu.edu.ph

ABSTRACT

Facial expressions have been proven to be universal or similar across cultures. However, this theory has been challenged by many researchers as there is evidence that suggests that facial expressions may differ from culture to culture. There are research works on Filipino embodied agents, however, none has been able to display emotions based on a corpus of Filipino facial expressions. This research presents an embodied agent that can express Filipino basic emotions such as happiness, sadness, fear, disgust, surprise and anger through facial expression. A corpus of Filipino facial expression is built which is used in creating the embodied agent.

KEYWORDS

Filipino Facial Expressions, Emotion Recognition, Embodied Agent

1 INTRODUCTION

Facial expressions are one of the many ways we humans communicate our inner emotions with one another. Facial expressions can be used to convey messages and they are also a two-way form of communication [1].

In the 1970s, facial expressions were found to be universal, meaning they were shown in the same manner given the same context such that people from the preliterate cultures were able to display similar or matching expressions to their Western counterparts, without external influence from the media or Western culture [2] though there were still some differences in the level of intensities that were identified by different cultures [3]. However, in 2012, facial expressions were found to be not universal and in fact, varies from culture to culture [4] [5].

Embodied agents are intelligent agents with an accurate artificial representation or depiction of a human being that is given a specific task. In many cases, these are used in virtual training environments, story telling for young children, etc. A Filipino embodied agent based on the image and likeness of Dr. Jose Rizal, the Philippine National hero, has already been made though it is unable to depict Filipino facial expressions, due to there being no data of such available yet [6].

Due to the theory that expressions are not universal, there is now a need to create a corpus of emotional facial expressions based on each distinct culture to capture their expressions [5]. There has been a study on the emotional facial expressions of Filipinos. However, the study focused on the individual action units as well as hand gestures [7]. A corpus of Filipino emotions also exists, though the expressions that were captured were not genuine and were too highly exaggerated [8]. The corpus was further developed but it is still unclear if the facial expressions obtained are indeed genuine or acted out as the corpus was developed based on data obtained from a reality TV show [9]. As mentioned previously, there is a Filipino-based embodied agent that has been created, however, it does not express Filipino facial expressions grounded on any corpus or data.

2 RELATED WORKS

FACIAL EXPRESSIONS

Facial expressions refer to motions of certain important muscles on a person's face to express an emotion or feeling. The face expresses emotions sooner than people verbalize or even perceive their feelings [10]. These expressions are either reactionary, which means that they occur on certain stimuli, caused, or voluntary. Voluntary facial expressions are often socially conditioned and follow a cortical route in the brain. Conversely, involuntary facial expressions are believed to be innate and follow a sub-cortical route in the brain. In social interactions, facial expressions can serve as communicative signals when humans talk to one another. Facial expressions can even be considered as public information. Such expressions are also cues that are abruptly produced and can be useful to observers [1].

The question of whether or not these expressions are present across the globe has been tackled by many researchers. In the 1970s, facial expressions were thought to be universal across several cultures as people from the various cultural groups were able to recognize and interpret similar emotions. These emotions being happiness, sadness, anger, fear, surprise and disgust - the 6 basic emotions found to be common across societies [11]. It was later on deduced that certain facial expressions coincide with certain emotions, as both people from literate and preliterate cultures were able to distinguish and identify the emotions of one another. However, the difference comes from the manner in which the emotion and subsequent expression is elicited [2].

Further research found that facial expressions were not universal as it was discovered that different cultures had different ways of expressing their emotions such as those from Asian cultures were said to suppress their emotional facial expressions while Europeans were said to be more expressive [5] [12]. With each culture having distinct emotional facial expressions and perceived emotional intensities, gathering facial data from each will be necessary to create an embodied agent that can depict accurate cultural facial expressions. As such, this research attempts to build an embodied agent that expresses Filipino basic emotions purely based on data collected from Filipino facial expressions.

Facial expressions come naturally depending on the emotion the person is feeling. Although that is usually the case, there are methods that can stimulate emotions so that the facial expression seems natural and not come out awkwardly [13]. Showing the person videos, images, making them listen to sounds/voices, music, and making them recall are valid methods of making a person elicit a genuine emotion. However, videos may stimulate multiple emotions at once and the only way to find out which emotion was the most prominent one throughout the video is by getting the intensities of the emotions felt by the viewer [14]. These methods for emotion elicitation are adapted in this study to generate genuine and authentic facial expressions from our test subjects.

There are various intensities to emotions which are found to be linearly proportional to the facial expression that the person is showing. However, this may not be applicable to both genders since females have been found to be more emotionally expressive than males [15]. For example, a female's emotion may be assumed to be at a lower intensity than what they were portraying it to be. With these findings, this research has created a corpus of emotional facial expressions, as it will be composed of various intensities of the basic emotions alongside the neutral facial expression.

FACIAL EXPRESSION RECOGNITION

Though humans can naturally recognize emotions, computers require specific software and methods to recognize human emotion through expressions. Several methods for facial recognition exist, and have been used in previous studies to recognize and classify facial expressions.

A comprehensive system that could differentiate all possible variations of facial movements which are represented as action units was developed and is known as the Facial Action Coding System

(FACS) [18]. Action units (AU) refer to significant points or the movement of specific muscles on the face, each of which causes distinct facial distortions. A total of 44 action units were defined and are encoded within the FACS. Each action unit correspond to an activity in a distinct muscle or muscle group [19]. The FACS splits the face into two sections, the upper and lower section which are further subdivided into action units [20]. The FACS can capture almost any action that can be performed by the human face, which would also allow us to capture micro-expressions [21]. The resulting system allows researchers to form emotional facial expressions through the combination of certain action units. For example, the combination of the action units 6 and 12 would result in the facial expression for happiness or joy while the combination of action units 1, 4 and 15 would result in the facial expression for sadness. However, there are a few combinations of action units that would never occur at the same time. These were labelled as antagonistic action units, an example would be that the lips could not be closed and opened in one expression [18].

EMBODIED AGENTS

Embodied agents are intelligent agents made to be interacted within a certain environment. Embodied agents may take on the form of anything ranging from a paperclip to a human. However, people prefer embodied agents with human forms over non-human agents [26]. Since we will be creating an embodied agent based on Filipinos, these studies will help us further understand how we should create our embodied agent in order for it to be an agent that users will be willing to engage with.

An Embodied Conversational Agent (ECA) modeled after Dr. Jose Rizal, the Philippine National hero has been developed [6] . The ECA's purpose was to teach English to elementary students, as it is also claimed that an embodied agent with the same cultural aspects as the student would allow it to be more effective and help sustain engagement. The ECA would be able to react accordingly based on the situation, as it was able to determine the emotion of the student by analyzing their speech and facial expressions. For example, if a student is able to answer a question correctly, the ECA would reply "Ang galing mo!" to give positive feedback to them. The ECA could also encourage students who were answering incorrectly or who seemed discouraged, by supporting the student through several levels of feedback. These involved repeating the word slowly, or repeating the word and emphasizing certain syllables to encourage the student to try again [6].

Greta is one of the first humanoid embodied agents that has a facial model capable of realistic expressions through features such as wrinkles and furrows. Greta was made using a pseudomuscular approach, meaning when a facial point is moved, the surrounding areas will also get deformed, forming wrinkles or furrows by using a bump mapping technique [28], to give her a realistic feel. An example would be when Greta furrows her eyebrows, wrinkles between her eyebrows will form [29].

3 METHODOLOGY

To create the embodied agent that can express authentic Filipino emotions, the researchers have gathered data on the facial expressions of Filipino citizens. As seen in Figure 1, the process of creating the embodied agent begins with the building of the corpus of emotional facial expressions. Annotators will tag time frames with an emotion and intensity. Frames of the start, middle, and end of the time frames will be extracted. Frame normalization and feature extraction will be done to build the corpus. Cartesian coordinates of the facial points will be used in model building and creating the embodied agent.

To build the corpus, fourteen (14) Filipino adults were asked to watch video clips with the attempt to elicit the six basic emotions while their facial expressions are captured by a camera. From the facial videos, relevant facial points were extracted

and normalized. The normalized data was used to create the corpus of emotional facial expressions which was then used to build a Filipino Embodied Agent capable of expressing Paul Ekman's basic emotions.

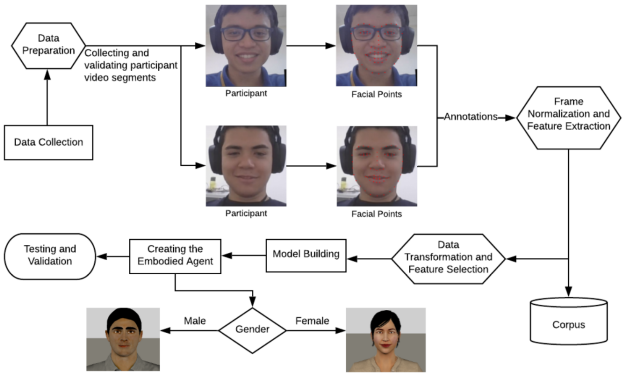


Figure 1: Diagram of Corpus and Embodied Agent creation process.

DATA COLLECTION

A total of fourteen (14) Filipino citizens aged 20-22 years old, participated in this study. The participants were asked to watch video clips with the attempt to elicit the six basic emotions while their facial expressions were being captured by the computer camera.

Figure 2 presents the typical set-up of the data collection. Each participant is asked to sit up straight approximately 1-3 meters away from a laptop with minimal movement and not obstruct their face in any way. Screen angle will be adjusted to ensure that the entirety of the participant's face is captured. The angle or orientation of the laptop camera is adjusted based on the participant's height to ensure that the participant's whole face is captured and to minimize chances of blurring. The lighting and the positioning of the table are well controlled to minimize distraction during the experiment. The facial expressions of the participants while watching video clips were captured using the laptop camera that produces an output with a resolution of 1280 x 720 pixels.

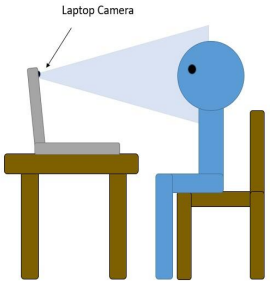


Figure 2: Setup of data collection.

After watching each video clip, the participant is asked to label the emotion and the intensity of which.

DATA PROCESSING AND DATA NORMALIZATION

Two (2) annotators with high emotional intelligence also provided annotation of the facial videos of the participants. The observed emotion and its intensity as well as the exact time in the video where the emotion was observed, were indicated by the annotators.

From the identified video segment, 3 frames were extracted - one is chosen in the first few seconds, one in the middle or the "peak", and another one towards the end of the segment. These frames were normalized by scaling and rotating the images so that all the faces would have the same size and orientation.

The normalization was based on detecting the position of the eyes, and the rotation is done by rotating the face until the eyes fall parallel to the x-axis. To calculate how much rotation is to make it a straight line, an arctan function is used. The center of rotation is then computed for by finding the midpoint between the eyes. The scaling is done at the same time, by using the distance between the eyes as a reference and using the rotation matrix function of OpenCV [42]. When the rotation matrix is applied onto the image, it is rotated and

scaled based on the values inside the rotation matrix (rotation angle, midpoint, scale ratio, etc.).

FEATURE EXTRACTION AND CORPUS BUILDING.

A total of 68 facial points were extracted from each frame and exported into a comma-separated values (csv) file. Open source libraries to perform such were used since they have the necessary functions in extracting facial landmarks.

USE OF DECISION TREE IN BUILDING EMBODIED AGENT

The normalized and pre-processed data was used in building the embodied agent. The data is used to train a model using a Decision tree (DT) algorithm using 10-fold cross validation. DT is selected since its model produces a tree that can provide insightful information like features that are found to be important in the classification. They are easy to understand and interpret due to the visualization of the tree. As such, facial points and their threshold values that are found to be significant in expressing basic emotions were identified from the DT.

Two sets of data were utilized. The first set of data has produced thresholds of the Cartesian coordinates in classifying the emotion being expressed. The second set of data has produced thresholds of the Cartesian coordinates in classifying the intensity of the emotion being expressed.

CREATING THE EMBODIED AGENT

SmartBody [43], an open-source character animation is used to create the male and female embodied agents. The 3D models for each agent are modified versions of 3D models already present in the SmartBody software. Modifications to the models were made primarily to make the embodied agents have the physical characteristics of a Filipino, such as a brown skin tone and black hair.

The embodied agent's facial expressions, each with three (3) different intensities, were then created. The intensity of the emotions ranges from one (1) to three (3), where 1 is low, 2 is moderate, and 3 is high. The embodied agent's initial facial expression for each emotion's lowest intensity (1) were based on descriptions of each emotion, specifically action units which were attributed with certain emotions [44]. However, since SmartBody only allows for manipulation of a set number of action units rather than facial points, adjustments to the face were made by identifying which action units present in SmartBody affected each facial point.

After an initial facial expression was manually adjusted on the embodied agent, the embodied agent's face was captured and run through the same software used to normalize and detect facial points. The facial point coordinates generated were then compared with the decision tree for the specific emotion. If the facial expression did not satisfy the thresholds for its targeted intensity, the conflicting facial point is adjusted and the process is repeated. This adjustment is based on the facial point threshold value generated by the decision tree. For example, if the x or y value of a specific facial point is above or below the decision tree threshold, then the action unit handling this facial point is adjusted accordingly. This was done until the facial expression satisfied the thresholds produced by the decision tree for the specific intensity. The facial point values of the succeeding intensities were then based on the previous intensities' facial point values. This was repeated until all 3 intensities satisfied the threshold given by the decision tree. This process was done for all 6 emotions. Figure 3 shows a flowchart describing how the data from the decision tree was used to create each facial expression on the embodied agents.

SmartBody was used to create the user interface which allows the user to see which facial expression and its intensity is currently being shown. Changing facial expressions and its intensities are also done through the same software through assigned keystrokes. Figure 4 presents the male and female agents expressing happiness.

TESTING AND VALIDATION

To validate whether the embodied agent expresses an authentic Filipino basic emotion, a Filipino registered psychologist was invited to evaluate how well the expression of the agent for 6 basic emotions with varying intensities is. Table 1 presents some of the observations of the psychologist.

In sum, the evaluation shows that the facial expressions are not strong enough to convey emotion, notably the eye and mouth expressions. Only a few facial expressions were able to convey their respective emotion and intensity clearly.

4 CONCLUSION

The goal of this research is to create a corpus of Filipino emotional facial expressions, develop a system that can automatically capture facial landmarks, and an embodied agent that can express those emotions to aid in future projects involving Filipino education and psychology. The corpus is built from authentic facial expressions, in the form of the x and y coordinates of the 68 facial points that were cleaned, normalized and filtered. The system that automatically captures facial landmarks was developed using existing Python libraries such as dlib and OpenCV and able to save data automatically to a readable format. The embodied agent we developed has the facial features and appearance of a Filipino, as well as the ability to express the 6 basic emotions based on the data gathered from the corpus. A character animation software, Smartbody, is used to create male and female embodied agents that can express Filipino happiness, sadness, fear, disgust, surprise and anger in 3 different intensities.

Decision tree is used to determine which facial points are found to be significant in expressing a particular emotion. These points are used as basis for adjusting the facial expression of the embodied agent. Future work may include extracting all the features that have values more than the threshold even if they do not appear in a decision tree model. Having 30 or more participants would significantly help the research in the long run because there would be more instances of emotions. This would immensely help in balancing the data set, as well as model building. Moreover, a more meaningful data may be gathered if the selected participants are more emotionally expressive.

The ECA developed in this study can only express human emotions based on data from Filipinos. Future work may improve the embodied agent by intelligently speak a language, in particular the Filipino language. This can further help in creating more opportunities for it to be used in applications made for medicine, and education, particularly for children.

REFERENCES

[1] C. Frith, "Role of facial expressions in social interactions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3453–3458, 2009.

[2] P. Ekman and W. V. Friesen, "Constants Across Cultures in the Face and Emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[3] P. Ekman, W. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. LeCompte, T. Pitcairn, and P. Ricci-Bitti, "Universals and cultural differences in the judgements of facial expressions of emotion," *Journal of personality and social psychology*, vol. 53, no. 4, pp. 712–717, 1987.

[4] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[5] W. Sato, S. Hyniewska, K. Minemoto, and S. Yoshikawa, "Facial expressions of basic emotions in Japanese laypeople," *Frontiers in Psychology*, vol. 10, pp. 1–11, 2019.

[6] J. Cu, J. Del Barrio, K. Villena, G. Ablaza, C. Musni, and M. Suarez, "Framework for an Empathic Filipino Embodied Conversational Agent for an Intelligent Tutoring System," *Philippine Computing Journal*, vol. 6, pp. 48–52, 2011.

[7] T. Tiam-Lee and K. Sumi, "Analyzing facial expressions and hand gestures in filipino students 'programming sessions," *Systems Man and Cybernetics*, pp. 1465–1470, 2018.

[8] J. Cu, M. Dy, I. Espinosa, P. Go, and C. Mendez, "Multimodal emotion recognition using a spontaneous filipino emotion database," 2010.

[9] J. Cu, K. Y. Solomon, M. T. Suarez, and M. S. Maria, "A multimodal emotion corpus for filipino and its uses," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 135–142, 2013.

[10] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1–19, 2001.

[11] P. Ekman, "Universal Facial Expressions of Emotions," *California Mental Health Research Digest*, vol. 8, no. 4, pp. 151–158, 1970.

[12] A. Murata, J. Moser, and S. Kitayama, "Culture shapes electrocortical responses during emotion suppression," *Social Cognitive and Affective Neuroscience*, vol. 8, no. 5, pp. 595–601, 2012.

[13] K. S. Quigley, K. A. Lindquist, and L. F. Barrett, "Inducing and measuring emotion and affect: Tips, tricks, and secrets." in *Handbook*

of research methods in social and personality psychology, H. Reis and C. Judd, Eds. Cambridge University Press, 2014, pp. 220–252.

[14] L. Rumpa, A. Wibawa, M. Heri Purnomo, and H. Tulak, "Validating video stimulus for eliciting human emotion: A preliminary study for ehealth monitoring system," in *2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*. IEEE, 2015, pp. 208–213.

[15] U. Hess, S. Blairy, and R. Kleck, "The Intensity of Emotional Facial Expressions and Decoding Accuracy," *Journal of Nonverbal Behavior*, vol. 21, no. 4, p. 241–257, 1997.

[16] S. Buisine, S. Abrilian, R. Niewiadomski, J.-C. Martin, L. Devillers, and C. Pelachaud, "Perception of blended emotions: From video corpus to expressive agent," in *International Workshop on Intelligent Virtual Agents*. Springer, 2006, pp. 93–106.

[17] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[18] P. Ekman and W. Friesen, "Measuring facial movement," *Environmental Psychology and Nonverbal Behavior*, vol. 1, pp. 56–75, 1976.

[19] M. S. Bartlett, P. A. Viola, T. J. Sejnowski, B. A. Golomb, J. Larsen, J. C. Hager, and P. Ekman, "Classifying facial action," pp. 823–829, 1996.

[20] J. J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Automated facial expression recognition based on facs action units," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 390–395.

[21] S. Kaiser and T. Wehrle, "Automated coding of facial behavior in human-computer interactions with facs," *Journal of Nonverbal Behavior*, vol. 16, no. 2, pp. 67–84, 1992.

[22] S. Gong, P. McOwan, and C. Shan, "Facial expression recognition based on local binary patterns: a comprehensive study," *Image and Vision Computing*, vol. 27, pp. 803–816, 2009.

[23] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE transactions on image processing*, vol. 16, no. 1, pp. 172–187, 2006.

[24] O. Deniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using Histograms of Oriented Gradients," *Pattern Recognition Letters*, vol. 32, pp. 1598–1603, 2011.

[25] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous¨ micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE, 2013, pp. 1–6.

[26] J. Zimmerman, E. Ayoob, J. Forlizzi, and M. McQuaid, "Putting a Face on Embodied Interface Agents," *Carnegie Mellon University*, 2005.

[27] C. Pelachaud, "Modelling Multimodal Expression of Emotion in a Virtual Agent," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 364, no. 1535, pp. 3539–48, 2009.

[28] J. F. Blinn, "Simulation of Wrinkled Surfaces," *SIGGRAPH Comput. Graph.*, vol. 12, no. 3, pp. 286–292, 1978.

[29] F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. De Carolis, "From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 81 – 118, 2003.

[30] K. Balci, E. Not, M. Zancanaro, and F. Pianesi, "Xface open source project and smil-agent scripting language for creating and animating embodied conversational agents," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 1013–1016.

[31] L. Alam and M. M. Hoque, "A text-based chat system embodied with an expressive agent," *Advances in Human-Computer Interaction*, vol. 2017, 2017.

[32] MakeHuman Community, "Makehuman," 2019. [Online]. Available: http://www.makehumancommunity.org

[33] Blender Foundation, "Blender - a 3d modelling and rendering package," Blender Foundation, Stichting Blender Foundation, Amsterdam, 2019. [Online]. Available: http://www.blender.org

[34] I. Mlakar, Z. Kaciˇ c, M. Borko, and M. Rojc, "A novel unity-basedˇ realizer for the realization of conversational behavior on embodied conversational agents," *International Journal of Computers*, vol. 2, pp. 205–213, 2017.

[35] Unity Technologies, "Unity," 2019. [Online]. Available: https://unity.com

[36] I. Revina and W. S. Emmanuel, "A survey on human facial expression recognition techniques," *Journal of King Saud University - Computer and Information Sciences*, 2018.

[37] J. Huang, S. Gutta, and H. Wechsler, "Detection of human faces using decision trees," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*. IEEE, 1996, pp. 248–252.

[38] P. Gupta. (2017) Decision trees in machine learning. [Online]. Available: https://towardsdatascience.com/decision-trees-inmachine-learning-641b9c4e8052

[39] R. Neath and M. Johnson, "Discrimination and Classification," in *International Encyclopedia of Education (Third Edition)*, third edition ed., P. Peterson, E. Baker, and B. McGaw, Eds. Elsevier, 2010, pp. 135 – 141.

[40] P. Thomas, "Perceptron learning for classification problems - impact of cost-sensitivity and outliers robustness," 01 2015, pp. 106–113.

[41] D. E. King, "dlib," 2019. [Online]. Available: http://dlib.net/

[42] Intel Corporation, Willow Garage, and Itseez, "OpenCV (Open Source Computer Vision Library)," 2019. [Online]. Available: https://opencv.org/

[43] Shapiro, A. (2011). Building a character animation system. In J. Allbeck &P. Faloutsos (Eds.),Motion in games(Vol. 7060, p. 98-109). SpringerBerlin / Heidelberg.

[44] Matsumoto, D., & Ekman, P. (2008). Facial expression analysis.Scholarpedia,3(5), 4237.

Table 1: Psychologist's Comments on the Embodied Agent's Facial Expressions Per Emotion per Gender

Emotion	Male	Female
Happiness	Wider smile; Squinting eyes do not convey happiness	Looks natural
Surprise	Lacking emotion: eyes and mouth	Intensity 3 looks closest to being surprised
Disgust	Improve expressions	Looks natural; Intensities 2 and 3 conveyed disgust
Sadness	Prolong time for sad face (lip gesture) to establish emotion	Only intensity 3 looks natural
Anger	Make the expression stronger, does not convey anger	Only intensity 3 looks natural
Fear	Improve lip expression; Eyes look okay	Presenting more worry than fear. Lip expression can be further improved.

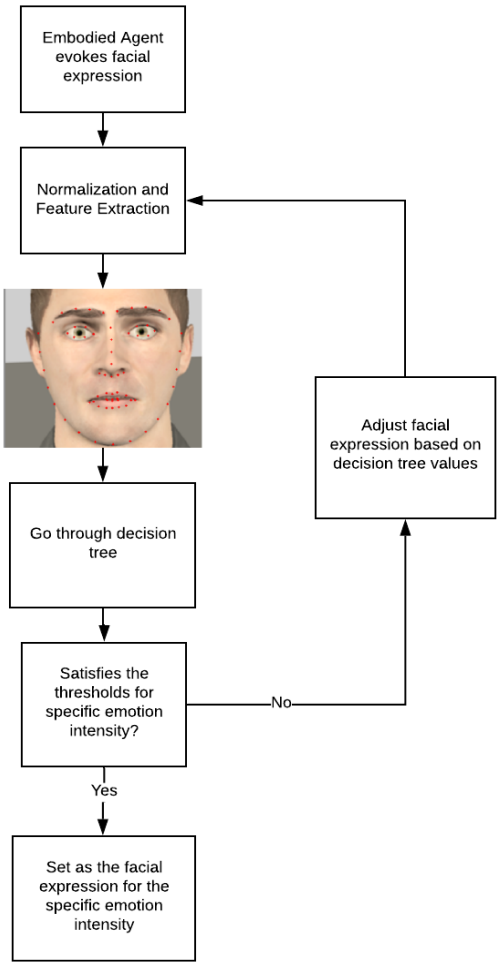


Figure 3: Flowchart on how the embodied agents facial expressions were adjusted based on decision tree data

Figure 4: Male and Female embodied agents expressing happiness with intensity 2.

CSP SIG on Information and Computing Education (SPICE)

SPICE Board Members, School Year 2018-2019

Jaime D.L. Caro
University of the Philippines, Diliman, Quezon City

Eric John G. Emberda
University of Immaculate Conception, Davao City

Gregg Victor D. Gabison
University of San Jose – Recoletos, Cebu City

Randy S. Gamboa
University of Southeastern Philippines, Davao City

Dave E. Marcial
Silliman University, Dumaguete City

Ramon L. Rodriguez
National University, City of Manila

Cherry Lyn Sta Romana
Cebu Institute of Technology – University, Cebu City

John Peter Abraham Ruero
Holy Angel University, Angeles City

Allan A. Sioson
Cobena Business Analytics and Strategy, Inc., BGC, Taguig City

COMPUTING SOCIETY OF THE PHILIPPINES

Board of Directors, School Year 2018-2019

President: **Rachel Edita O. Roxas, PhD**
National University, City of Manila

Vice President: **Randy S. Gamboa, PhD**
University of Southeastern Philippines, Davao City

Secretary: **Rafael A. Cabredo, PhD**
De La Salle University, Taft, City of Manila

Treasurer: **Nathaniel A. Oco**
National University, City of Manila

Asst. Treasurer: **Marrick C. Neri, PhD**
University of the Philippines, Diliman, Quezon City

Auditor: **Charibeth Cheng**
De La Salle University, Taft, City of Manila

Henry N. Adorna, PhD
University of the Philippines, Diliman, Quezon City

Proceso L. Fernandez, Jr., PhD
Ateneo de Manila University, Loyola Heights, Quezon City

Mary Jane Sabellano
University of San Carlos, Cebu City