Recognizing Non-manual Signals in Filipino Sign Language

Joanna Pauline Rivera, Clement Ong
De La Salle University
Taft Avenue, Manila 1004 Philippines
joanna_rivera@dlsu.edu.ph, clem.ong@delasalle.ph

Abstract

Filipino Sign Language (FSL) is a multi-modal language that is composed of manual signlas and non-manual signals. Very minimal research is done regarding non-manual signals (Martinez and Cabalfin, 2008) despite the fact that non-manual signals play a significant role in conversations as it can be mixed freely with manual signals (Cabalfin et al., 2012). For other Sign Languages, there have been numerous researches regarding non-manual; however, most of these focused on the semantic and/or lexical functions only. Research on facial expressions in sign language that convey emotions or feelings and degrees of adjectives is very minimal. In this research, an analysis and recognition of non-manual signals in Filipino Sign Language are performed. The non-manual signals included are Types of Sentences (i.e. Statement, Question, Exclamation), Degrees of Adjectives (i.e. Absence, Presence, High Presence), and Emotions (i.e. Happy, Sad, Fast-approaching danger, Stationary danger). The corpus was built with the help of the FSL Deaf Professors, and the 5 Deaf participants who signed 5 sentences for each of the types in front of Microsoft Kinect sensor. Genetic Algorithm is applied for the feature selection, while Artificial Neural Network and Support Vector Machine is applied for classification.

Keywords: non-manual signal, Filipino Sign Language, Kinect, machine learning

1. Introduction

Filipino Sign Language (FSL) is a communication medium among the Deaf in the Philippines. It originally rooted from American Sign Language (ASL) (Hurlbut, 2008) but soon developed and became a separate language from ASL as Filipino Deaf used it in communication at school passing Filipino signs that emerged naturally through generations and adding new signs, especially those that are related to technology (Apurado and Agravante, 2006).

FSL has five components: hand shape, location, palm orientation, movement, and non-manual signal (Philippine Deaf Research Center and Philippine Federation of the Deaf, 2004).

1. **Hand shape** is the arrangement of the fingers and their joints.
2. **Location** is the place where the hand/s is/are.
3. **Palm orientation** is where the palm is facing.
4. **Movement** is the change in hand shape and/or path of the hands.
5. **Non-manual signals** are facial expressions and/or movement of the other parts of the body that goes with the signing.

Several researches in the computing field regarding FSL have been conducted despite the only recent linguistic researches about it. However, most studies in Filipino Sign Language (FSL) focuses on recognizing manual signals only. Very minimal research was done regarding non-manual signals and its integration with the manual signals (Martinez and Cabalfin, 2008) even though non-manual signals play a significant role in conversations as it can be mixed freely with manual signals (Cabalfin et al., 2012).

A common misconception in Sign Language Recognition Systems is approaching the problem through Gesture Recognition (GR) alone (Cooper et al., 2011). Sign language is a multi-modal language that has two components: manual and non-manual signals. Manual signals are hand gestures, positions and shapes which convey lexical information. Non-manual signals are facial expressions, head movements, and upper body posture and movements which express syntactic and semantic information. These signals usually co-occur with manual signals often changing its meaning (Nguyen and Ranganath, 2008). Hence, SLR systems or techniques would not yield effective results without the non-manual signals (Von Agris et al., 2008).

As researchers realize the significance of non-manual signals, some studies focusing on facial expression recognition in Sign Languages were conducted (Von Agris et al., 2008). These studies focus more on Grammatical Facial Expressions (GFE) that convey semantic functions such as WH-question, Topic and Assertion.

Some of the studies that focus on recognizing facial expressions on Sign Language include: Grammatical Facial Expressions Recognition with Machine Learning (de Almeida Freitas et al., 2014), Recognition of Non-manual Markers in American Sign Language (Metaxas et al., 2012), and Spatial and Temporal Pyramids for Grammatical Expression Recognition of American Sign Language (Michael et al., 2009). Most of these studies differ in the data representation, and the machine learning technique. One similarity among them is that they all focused on GFEs that convey semantic functions such as WH-question, Topic and Assertion. On the other hand, facial expressions in sign languages are not limited to those semantic functions. For instance, facial expressions in ASL are used to convey degrees of adjectives (e.g. color intensity), adverbial information (e.g. carelessly) and emotions as well (Martinez and Cabalfin, 2008).

Similarly, non-manual signals in FSL is used to convey lexical information, types of sentences (what is referred to as semantic information in other studies), degrees of adjec-

tives, and emotions.

This study aims to recognize and analyze non-manual signals in FSL using the Microsoft Kinect and Machine Learning. The rest of the paper is organized as follows. In section 2, the process of building the corpus with the help of FSL signers is described. In section 3, the feature extraction method is discussed. In section 4, the Machine Learning techniques applied are enumerated. In section 5, the results and analysis for each non-manual signal category are explained. Lastly, the research is concluded and the recommendations are listed in section 6.

2. Corpus Building

2.1. Interviews with FSL Deaf Professors

To understand more about FSL, an interview was conducted with Ms. Maria Elena Lozada, a Deaf professor from the School of Deaf Education and Applied Studies at De La Salle - College of Saint Benilde. According to her, some non-manual signals in FSL are part of its lexicon such as "thin" and "fat". Others are used to differentiate sentences with the same signs but different semantics. For instance, the statement "John likes Mary" is signed the same way as the question "Does John like Mary?". The two sentences are differentiated using different non-manual signals. There are also non-manual signals that are used to convey the degrees of adjectives (e.g. "angry" and "very angry"). Lastly, as Filipinos were born naturally expressive, non-manual signals are mostly used to convey emotions or feelings (Lozada, 2016).

Another interview was conducted with Mr. Rey Alfred Lee, a Deaf professor from Filipino Sign Language Learning Program, Center Education Assessment for the Deaf, and School of Deaf Education and Applied Studies at De La Salle - College of Saint Benilde. From the interview, it was concluded that FSL uses Affective Facial Expressions (AFE) and Grammatical Facial Expressions (GFE), similar to other Sign Languages. AFEs in FSL are used to show emotions and feelings, while GFEs are used to convey lexical information, types of sentences, and degrees of adjectives. The types of sentences are further subdivided into question, statement, and exclamation, while the basic degrees of adjectives are subdivided into absence, presence, and high presence (Lee, 2016). In other researches, it is stated that GFE differ from AFE in terms of the facial muscles used (McCullough and Emmorey, 2009) and its behavior, such as form and duration (Muller, 2014).

Both professors are asked about the importance of non-manual signals in FSL. According to Ms. Lozada, although it is possible to use FSL without facial expressions, it would be difficult to carry out a conversation especially when telling a story which involves facial expressions to convey types of sentences, emotions, and degrees of adjectives (Lozada, 2016). According to Mr. Lee, being able to recognize the degrees of adjectives, specifically feelings, and emotions can also help the medical doctors and psychologists in determining the amount of pain that the signer feels and the emotional state of the patient (Lee, 2016).

In line with this, this research focuses on non-manual signals in FSL that convey Types of Sentences (i.e. statement, question, and exclamation), Degrees of Adjectives (i.e. absence, presence, and high presence), and Emotions.

The emotions considered are the four basic emotions (i.e. happy, sad, fast-approaching danger and stationary danger) (Jack et al., 2014). In their work, they have shown that there are only four basic emotions which were only discriminated into six (i.e. happy, sad, fear, surprise, disgust and anger) as time passed by. Fear and surprise can be combined as they both belong in the fast-approaching danger, while disgust and anger both belong in the stationary danger.

Table 1 shows a summary of the types of Facial Expressions in FSL that were used for this research. These types were used as labels by the FSL annotator.

Affective Facial Expressions	Emotions	Happy
		Sad
		Stationary danger
		Fast-approaching danger
Grammatical Facial Expressions	Types of Sentences	Statement
		Question
		Exclamation
	Degrees of Adjectives	Absence
		Presence
		High Presence

Table 1: Categories of Facial Expressions in Filipino Sign Language

2.2. Data Collection with FSL Signers

There are already existing corpus for FSL. However, these are built to focus on the manual signals data. The different types of non-manual signals may not be represented well on these corpus.

Thus, data collection is performed using Microsoft Kinect for Windows v2.0 sensor (Kinect sensor) (Microsoft, 2016). The Kinect sensor has a depth sensor, a color camera, an infrared (IR) emitter, and a microphone array that allow tracking of the location, movement, and voice of people (Zhang, 2012).

The 3D videos are collected from 5 participants, 20-24 year old third-year FSL students. Two of them are male while three of them are female. Their learning and actual experience in FSL is approximately 2-4 years. With regards to facial expressions in FSL, most of them have learned it in school about 1-3 months, while some of them have been using it for 1-3 years.

5 sentences for each type of facial expression, a total of 50 sentences, were signed by each participant. To assure that all samples are appropriate for its corresponding type, these sentences were formulated with the guidance of the FSL Deaf professor of the participants, Ms. Lozada. Refer to Table 2 for the complete list of sentences used for this study.

Type	Sentences
Happy	1. Thank you. 2. The trip is exciting. 3. The show is amazing. 4. I am proud of you! 5. Our team won!
Stationary Danger	1. I hate you! 2. You are disgusting! 3. I don't like you! 4. You are so slow! 5. Stay away from me!
Fast-approaching Danger	1. I am scared. 2. I am nervous. 3. I am worried. 4. I saw a ghost. 5. I am shocked!
Sad	1. I am sorry. 2. My dog died. 3. I am alone. 4. I am heartbroken. 5. I failed the exam.
Question	1. Does John like Mary? 2. Are you sick? 3. Is it new year? 4. How are you? 5. How old are you?
Statement	1. John likes Mary. 2. You are sick. 3. It is new year. 4. I am fine. 5. I am 12 years old.
Exclamation	1. John likes Mary! 2. You are sick! 3. Happy new year! 4. Good morning! 5. Good noon!
Absence	1. My head is not painful. 2. I do not like you. 3. I am not tired. 4. You are not slow. 5. This is not hard.
Presence	1. My head is painful. 2. I like you. 3. I am tired. 4. You are slow. 5. This is hard.
High Presence	1. My head is very painful. 2. I like you very much. 3. I am so tired. 4. You are so slow. 5. This is very hard.

Table 2: Sample Sentences for each of the types of Non-manual Signals in FSL

2.3. Data Annotation with FSL expert

Supposedly, the annotation label of each sentence is their intended type since the facial expressions are acted, see Table 2 for the intended type for each of the sentences. However, initial experiments show very poor performances reaching the highest accuracy of 26% using Artificial Neural Network (ANN). Looking at the confusion matrix shown in Table 3, it can be deduced that the classes are very confused with each other, meaning there are similarities between them.

true=	a	b	c	d	e	f	g	h	i	j
pred. a	**5**	4	5	2	3	1	6	1	0	0
pred. b	4	**4**	3	2	1	1	0	2	1	0
pred. c	4	5	**7**	0	3	2	1	2	3	1
pred. d	0	2	1	**7**	3	4	0	2	4	2
pred. e	1	0	1	2	**4**	6	4	1	1	4
pred. f	3	1	2	2	4	**2**	1	2	0	2
pred. g	1	1	1	4	2	1	**5**	2	1	3
pred. h	5	5	2	1	2	3	3	**10**	1	2
pred. i	1	3	3	3	1	1	2	1	**12**	2
pred. j	1	0	0	2	2	4	3	2	2	**9**

Table 3: Confusion matrix during initial experiment using ANN where: a=question, b=statement, c=Exclamation, d=absence, e=presence, f=high presence, g=stationary, h=fast, i=happy, and j=sad

With a consultation with an FSL annotation expert, co-occurrences of the different classes in a sample are discovered. As a result, there is a maximum of three labels in an instance. For example, the facial expression for specific sign/s can be a combination of question (one of the types of sentences), presence (one of the degrees of adjectives), and sad (one of the emotions). Thus, individual experiments were conducted for Types of Sentences, Degrees of Adjectives and Emotions, each applying the classification techniques.

3. Feature Extraction

Color images, depth images, audio input, and skeletal data from Kinect sensor are processed with the help of Microsoft Kinect for Windows Software Development Kit 2.0 (Kinect SDK) (Microsoft, 2016) to extract the features.

The face orientation, Shape Units (SU), and Animation Units (AU) are used as features for this study as most international research works have concluded that the eyes, eyebrows, mouth, nose, and head pose must be well-represented to achieve effective recognition. The face orientation is the computed center of the head which is used to calculate the angle rotations of the head with respect to the optical center of the camera of Kinect sensor (i.e. pitch, yaw, and roll). The SUs are the weights that indicate the differences between the shape of the face tracked and the

average shape of a person which is derived using the Active Appearance Models (AMM) of (Smolyanskiy et al., 2014). These SUs are used to indicate the neutral shape of the face which is derived from the first few frames. The AUs are the deltas in the facial features of the face tracked from the neutral shape.

In summary, the 20 features used are the pitch, yaw and roll angles, and the seventeen AUs shown in Table 4. Most of the values range from 0 to 1. Negative minimum value indicates delta on the opposite direction. For example, if the delta for EyebrowLowerer is -1, the eyebrows are raised instead of lowered.

Movement	Min Value	Max Value
Pitch		
Yaw		
Roll		
JawOpen	0	+1
LipPucker	0	+1
JawSlideRight	-1	+1
LipStretcherRight	0	+1
LipStretcherLeft	0	+1
LipCornerPullerRight	0	+1
LipCornerPullerLeft	0	+1
LipCornerDepressorLeft	0	+1
LipCornerDepressorRight	0	+1
LeftCheekPuff	0	+1
RightCheekPuff	0	+1
LeftEyeClosed	0	+1
RightEyeClosed	0	+1
RighteyebrowLowerer	-1	+1
LefteyebrowLowerer	-1	+1
LowerlipDepressorLeft	0	+1
LowerlipDepressorRight	0	+1

Table 4: Face Orientation and Animation Units with the minimum and maximum weights

4. Machine Learning

Before the data has undergone classification, some pre-processing tasks are performed. Particularly, only samples based from peak facial expressions are selected since some frames between the neutral and peak facial expressions showed hand occlusions on the face. This also ensures that rising and falling facial expressions are excluded from the samples. Afterwards, uniform undersampling is applied since the data for Degrees of Adjectives is imbalanced. Normalization through z-transformation is also applied due to the different ranges of feature values.

Then, feature selection is applied to determine the most effective features for each category. The Wrapper Subset Evaluation dominated in terms of improving the accuracy; however, it is computationally expensive (Hall and Holmes, 2003). Thus, Genetic Algorithm is applied to reduce the amount of resources needed.

Some of the most commonly used classifiers in recent studies regarding Facial Expression Recognition in Sign Language are Artificial Neural Network (ANN), and Support Vector Machine (SVM) (Mao and Xue, 2011). ANN with a learning rate of 0.3, and SVM with a kernel type of radial basis function are applied for this study. Then, the validation technique used is k-fold Cross-Validation while the performance metrics are accuracy and kappa.

5. Experiments, Results and Analysis

Several experiments are conducted to analyze the types of non-manual signals in FSL. These experiments can be categorized into 3: Participants-based, Features-based, and Class-based.

The Participants-based experiments are subdivided into Partitioning by Participants and Partitioning by Sentences. In Partitioning by Participants setup, there are five folds for the validation phase. In each fold, there are four participants in the training set while there is one participant in the test set. In Partitioning by Sentences, there are a total of 10 folds for the validation phase. In each fold, 90% of the sentences are in the training set, while 10% are in the test set.

Using Participants-based experiments, findings indicate that there are not much differences on the performances for all categories between the Participant-based experiments as shown in Figure 1 and Figure 2. This suggests that introducing a new face would not have much impact on the classification. This is because AUs are deltas and not the exact points on the face. Thus, different facial structures would not matter that much as long as the participants are all expressive. In Sign Language, facial expressions are significant; thus, signers are usually expressive.

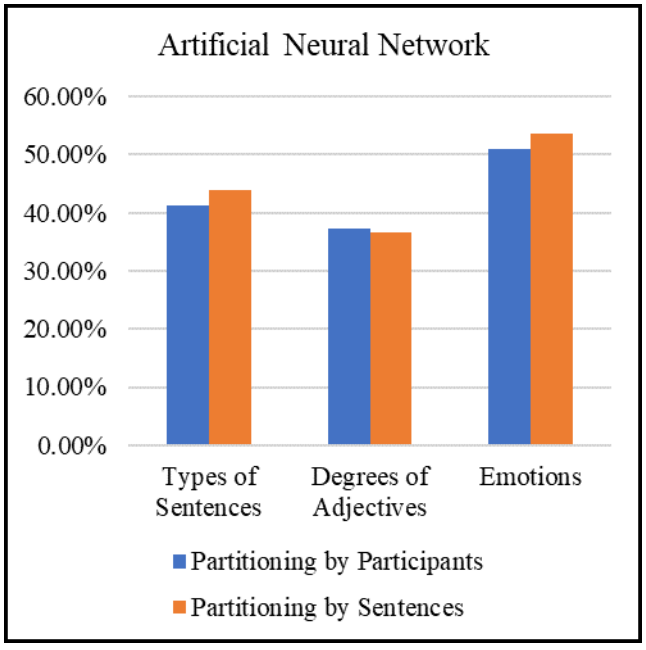

Figure 1: Comparison of performances using Partitioning-based setups for ANN

The Features-based experiments rely on adding classes from other categories as features. For example, the features for Degrees of Adjectives may include Fast-approaching danger, Stationary danger, Happy, and Sad. This is an attempt to represent the intensities for degrees of adjectives, and the co-occurrences of the different categories in one in-

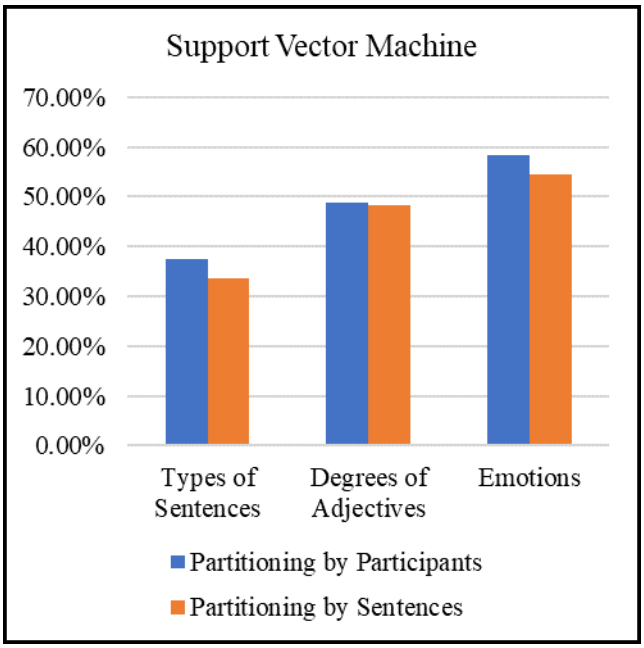

Figure 2: Comparison of performances using Partitioning-based setups for SVM

stance. The idea here is that the Degree of Adjective is Presence if the feature values are within the average given a co-occurring Emotion. It is High Presence if the feature values are higher than the average given a co-occurring Emotion.

Using Features-based experiments, findings indicate that adding classes from other categories as features are effective in representing the intensities, and the co-occurrences of the different categories in one instance reaching an increase of 17% to 30% recognition rate.

The Class-based experiments are subdivided into: One versus Many and One Class Removed. In One versus Many, one class is retained while the remaining classes are merged to form a new class. For example, the possible classes for Degrees of Adjectives are Presence or Not, Absence or Not, and High Presence or Not. In One Class Removed, one class is not included for each category during the experiments. For examples, the possible classes for Types of Sentences are Statement or Question, Statement or Exclamation, and Question or Exclamation.

Using Class-based experiments, highest performances for all categories are achieved which implies that some essential features for other classes are not represented. This is because motions and context are not represented which are significant for some of the classes based on the direct observation of the video data.

5.1. Types of Sentences

Distinction of Question and a confusion between Statement and Exclamation are observed in class-based experiments as shown in Table 5. It is also observed that Question has more distinction with Statement than with Exclamation.

Similar findings and further improvements were observed using Feature-based setups. Adding Emotions and Degrees of Adjectives as features resulted to the highest performances. Hence, Types of Sentences are highly affected by the co-occurrences of the other categories. Refer to Table 6 for the performances in accuracies and kappas.

Classes	ANN		SVM	
	Acc.	Kappa	Acc.	Kappa
Exclamation or Not	55.00%	0.220	53.33%	0.110
Question or Not	61.33%	0.231	62.33%	0.264
Statement or Not	63.33%	0.267	53.33%	0.067
Statement-Question	57.67%	0.199	60.00%	0.265
Statement-Exclamation	46.67%	-0.060	43.33%	-0.070
Question-Exclamation	60.00%	0.210	46.67%	-0.040

Table 5: Comparison of the performances of ANN, and SVM using Class-based setups for Types of Sentences

Setup	ANN		SVM	
	Acc.	Kappa	Acc.	Kappa
Emotions and Degrees	60.00%	0.409	58.00%	0.363
Emotions	40.00%	0.089	30.50%	-0.008
Degrees	43.00%	0.114	52.00%	0.277

Table 6: Comparison of the performances of ANN, and SVM using Feature-based setups for Types of Sentences

Since adding Emotions and Degrees of Adjectives as features lead to higher performances, Genetic Algorithm for feature selection is applied after performing this setup. Refer to Table 7 for the list of features selected along with the weights. Applying feature selection and adding Emotions and Degrees of Adjectives resulted to performances reaching accuracy of 76.00% and kappa of 0.619 using ANN. Similar behavior on the confusion matrices are also observed but the distinction between Statement and Exclamation improved.

From the observation of the videos, Question is mostly characterized by eyes, eyebrows, lip, and head movements which makes it distinct from the other classes. The eyes movements are captured by LeftEyeClosed and RightEyeClosed. The eyebrows movements are captured by RightEyebrowLowerer and LeftEyebrowLowerer. The lip movements (i.e. lip corners pulled downwards) are captured by LipCornerPullerLeft, LipCornerPullerRight, LipCornerDepressorRight, and LipCornerDepressorLeft. Lastly, the head movements are captured by pitch, yaw, and roll. Since all the distinct characteristics of Question are represented by the head rotation angles and seventeen AUs, Question always has the highest precisions and recalls.

On the other hand, Statement and Exclamation are always confused. Looking at the videos, these classes do not have much distinguishing features. Statement is like a neutral facial expression that changes based on the current Degree of Adjective or Emotion. When it is mixed with the other classes, it becomes similar to Exclamation.

Feature	Weight
emotions = n	1
emotions = happy	1
emotions = stationary	1
emotions = fast	1
degrees = absence	1
degrees = n	1
pitch	1
yaw	1
JawSlideRight	1
LipCornerPullerRight	1
LipCornerDepressorRight	1
RightcheekPuff	1
LefteyeClosed	1
RighteyebrowLowerer	1
LowerlipDepressorLeft	1

Table 7: Features Selected using Genetic Algorithm on Types of Sentences

Aside from this, only the difference in speed of the motions are observed. In this study, only the peak facial expressions are selected so the motions are not captured. Also, the head rotation angles and seventeen AUs alone cannot handle motions since these features are only concerned with the deltas between the current and the neutral facial expression.

5.2. Degrees of Adjectives

Distinction between Absence and High Presence is observed on One Class Removed of Class-based experiments as shown in Table 8. On the other hand, Presence cannot be distinguished from the rest of the classes as shown in One versus Many of Class-based experiments.

Setup	ANN		SVM	
	Acc.	Kappa	Acc.	Kappa
Presence or not	42.44%	-0.146	40.32%	-0.160
Absence or not	70.24%	0.372	70.00%	0.371
High Presence or not	59.09%	0.174	68.11%	0.354
High Presence-Absence	82.38%	0.631	87.14%	0.727
Presence-Absence	50.48%	0.019	52.86%	0.078
Presence-High Presence	46.67%	-0.059	56.97%	0.164

Table 8: Comparison of the performances of ANN, and SVM using Class-based setups for Degrees of Adjectives

Absence and High Presence are like polar opposites which is why they can easily be distinguished from each other. On the other hand, Presence is like a neutral facial expression similar to Statement. It becomes similar to the other classes when mixed with Emotions.

High Presence can be differentiated from Presence based on the intensity of the facial expression of the sentence. The intensity is represented through adding other classes as features. Results shown in Table 9 indicate that adding Emotions as features yield the best performances among Features-based experiments. This validates the idea that the Degree of Adjective is Presence if the feature values are within the average given a co-occurring Emotion. It is High Presence if the feature values are higher than the average given a co-occurring Emotion.

Setup	ANN		SVM	
	Acc.	Kappa	Acc.	Kappa
Emotions and Sentences	47.22%	0.218	54.78%	0.330
Emotions	51.22%	0.265	62.33%	0.435
Sentences	33.67%	0.014	48.56%	0.231

Table 9: Comparison of the performances of ANN, and SVM using Feature-based setups for Degrees of Adjectives

Absence can be differentiated from Presence by detecting the motion of head shaking. However, the head rotation angles and seventeen AUs only represent the delta between the neutral face and the peak facial expression. The motion yawing to the left or right can be captured by the AUs, but not the whole motion of head shake.

Genetic Algorithm is applied for feature selection after adding Emotions as features. Refer to Table 10 for the complete list with the weights. Without removing classes or merging classes, the highest accuracy reached 70.89% with a kappa of 0.562 using SVM.

Feature	Weight
yaw	1
JawOpen	1
LipStretcherRight	1
LipCornerPullerLeft	1
RighteyebrowLowerer	0.938899
LefteyeClosed	0.89875
LipCornerDepressorLeft	0.735377
LowerlipDepressorLeft	0.429755
LeftcheekPuff	0.413458
pitch	0.221545

Table 10: Features Selected using Genetic Algorithm on Degrees of Adjectives

5.3. Emotions

Results from Class-based experiments indicate that Happy and Fast-approaching danger are distinct from the other classes using One versus Many (i.e. Happy or Not, and Fast-approaching danger or Not), while Sad and Stationary danger are confused with each other as shown using One Class Removed (i.e. Happy or Sad or Stationary danger, and Fast-approaching danger or Sad or Stationary danger). Refer to Table 11 for the results of Class-based experiments.

In contrary to the effect of Features-based experiments on Types of Sentences and Degrees of Adjectives, adding classes from other categories as features did not have good effect on the performances. Refer to Table 12 for the results of Features-based experiments. It is shown that the highest

Setup	ANN		SVM	
	Acc.	Kappa	Acc.	Kappa
Fast or not	68.50%	0.367	73.50%	0.467
Happy or not	87.00%	0.739	85.33%	0.718
Sad or not	65.67%	0.291	74.67%	0.495
Stationary or not	74.00%	0.470	70.00%	0.398
Happy-Sad-Stationary	59.72%	0.398	55.00%	0.332
Fast-Sad-Stationary	55.71%	0.340	51.61%	0.278
Fast-Happy-Stationary	68.57%	0.529	79.46%	0.687
Fast-Happy-Sad	76.61%	0.642	78.93%	0.671

Table 11: Comparison of the performances of ANN, and SVM using Class-based setups for Emotions

accuracy of 67% is achieved when Sentences and Degrees are added as features. However, this is lower than when only Genetic Algorithm is applied resulting to 72.91% accuracy. This implies that Emotions is not affected by the co-occurrence of the other categories.

Setup	ANN		SVM	
	Acc.	Kappa	Acc.	Kappa
Sentences and Degrees	59.27%	0.454	67.00%	0.556
Degrees	54.73%	0.393	58.18%	0.443
Sentences	62.36%	0.492	65.00%	0.529

Table 12: Comparison of the performances of ANN, and SVM using Feature-based setups for Emotions

Happy and Fast-approaching danger have characteristics that make them distinct from the other classes. Happy is mostly characterized by a smiling face, while Fast-approaching danger are mostly characterized by eyes and/or mouth wide opened. On the other hand, the characteristics of Sad and Stationary danger are very similar which makes it difficult for the classifier to distinguish between the two. Stationary danger and sad are mostly characterized by a frowning face. A possible reason why the annotators can recognize it is their knowledge about the context shown by the gestures.

5.4. Summary

Without removing or merging classes, highest performances are achieved by adding classes from other categories as features and/or applying genetic algorithm for feature selection. For Types of Sentences, Emotions and Degrees of Adjectives are added as features and genetic algorithm is applied, reaching the highest accuracy of 76.00% and kappa of 0.619 using ANN. For Degrees of Adjectives, Emotions are added as features and genetic algorithm is applied, reaching the highest accuracy reached 70.89% with a kappa of 0.562 using SVM. For Emotions, genetic algorithm is applied reaching the highest performance of 72.91% accuracy and 0.639 kappa.

6. Conclusion and Recommendations

In this study, the different non-manual signals, specifically Types of Sentences, Degrees of Adjectives, and Emotions are recognized and analyzed towards aiding the communication between the Deaf community, and the medical doctors, psychologists, and other non-signers.

Based on the experiments conducted, AUs are effective in representing different facial structures of the signers, but motions, intensities, and co-occurrences of classes from other categories must also be well-represented. Representing the intensities and co-occurrences by adding classes from other categories as features yielded better performances. However, confusion matrices show that the representation of intensities must still be improved. In addition, the gesture data is important as it shows the context which can further help in distinguishing the facial expressions. As stated by the FSL experts and annotators, knowing the meaning of gestures help them annotate the facial expressions. Without the seeing the gestures it would be difficult for them to distinguish the facial expressions.

In line with the conclusion, in addition to head rotation angles and AUs, motions must be captured to represent the data better. In the studies of (Von Agris et al., 2008), (de Almeida Freitas et al., 2014), and (Nguyen and Ranganath, 2010), representations of motions through the inclusion of temporal information such as Sliding Window and Spatio-Temporal Pyramids improved their recognition rates. In the studies of (Metaxas et al., 2012), and (Nguyen and Ranganath, 2010), machine learning techniques that can handle temporal dynamics by making use of sequential data were applied such as Hidden Markov Model and Conditional Random Field respectively. Motions are not represented in this study since some frames were dropped due to hand occlusions. Based from the other works, removing the frames with hand occlusions is not the solution to the problem. Feature detection techniques that can handle occlusions must be applied such as Kanade-Lucas-Tomasi (KLT) with Bayesian Feedback Mechanism and Non-parametric Adaptive 2D-3D Tracking in the studies of (Metaxas et al., 2012) and (Nguyen and Ranganath, 2010) respectively, instead of AAM-based methods.

Aside from motions, intensities and co-occurrences must also be represented well. In this study, an attempt to represent these is adding classes from other categories as features. Significantly better performances were observed using this setup. However, this approach can still be improved. One way to recognize the intensities could be applying Convolutional Neural Network (CNN) for classification which is one of the state-of-the-art approach of deep learning for images.

Lastly, an integration of the gesture data can be explored as it contains the context that might be significant in distinguishing the facial expressions. As the annotator and other FSL experts have mentioned, annotating the data without seeing the gestures is possible but it would be difficult.

7. Acknowledgements

The authors would like to thank the anonymous participants from the De La Salle-College of St. Benilde who agreed to participate in the data gathering for the corpus building.

The authors would also like to express their deepest gratitude to Ms. Maria Elena Lozada for her support in the data collection, and Mr. Rey Alfred Lee and Mr. John Xander Baliza for imparting their knowledge on Filipino Sign Language.

8. Bibliographical References

Apurado, Y. S. and Agravante, R. L. (2006). The phonology and regional variation of Filipino Sign Language: considerations for language policy. In *9th Philippine Linguistic Congress*.

Cabalfin, E. P., Martinez, L. B., Guevara, R. C. L., and Naval, P. C. (2012). Filipino Sign Language recognition using manifold projection learning. In *TENCON 2012-2012 IEEE Region 10 Conference*, pages 1–5. IEEE.

Cooper, H., Holt, B., and Bowden, R. (2011). Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer.

de Almeida Freitas, F., Peres, S. M., de Moraes Lima, C. A., and Barbosa, F. V. (2014). Grammatical facial expressions recognition with machine learning. In *FLAIRS Conference*.

Hall, M. A. and Holmes, G. (2003). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Trans. on Knowl. and Data Eng.*, 15(6):1437–1447, November.

Hurlbut, H. M. (2008). Philippine signed languages survey: a rapid appraisal. *SIL Electronic Survey Reports*, 10:16.

Jack, R. E., Garrod, O. G., and Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192.

Lee, R. A. (2016). Facial expressions in Filipino Sign Language, March. Personal Communication.

Lozada, M. E. (2016). Facial expressions in Filipino Sign Language, February. Personal Communication.

Mao, X. and Xue, Y. (2011). Human Emotional Interaction.

Martinez, L. and Cabalfin, E. P. (2008). Sign language and computing in a developing country: a research roadmap for the next two decades in the Philippines. In *PACLIC*, pages 438–444.

McCullough, S. and Emmorey, K. (2009). Categorical perception of affective and linguistic facial expressions. *Cognition*, 110(2):208–221, February.

Metaxas, D. N., Liu, B., Yang, F., Yang, P., Michael, N., and Neidle, C. (2012). Recognition of nonmanual markers in American Sign Language (ASL) Using non-parametric adaptive 2d-3d face tracking. In *LREC*, pages 2414–2420. Citeseer.

Michael, N., Metaxas, D., and Neidle, C. (2009). Spatial and temporal pyramids for grammatical expression recognition of American Sign Language. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 75–82. ACM.

Microsoft. (2016). Kinect hardware requirements and sensor setup.

Muller, C. (2014). *Body - Language - Communication*. Walter de Gruyter GmbH & Co KG, October. Google-Books-ID: QAGTBgAAQBAJ.

Nguyen, T. D. and Ranganath, S. (2008). Tracking facial features under occlusions and recognizing facial expressions in sign language. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–7. IEEE.

Nguyen, T. D. and Ranganath, S. (2010). Recognizing Continuous Grammatical Marker Facial Gestures in Sign Language Video. In *Computer Vision - ACCV 2010*, pages 665–676. Springer, Berlin, Heidelberg, November.

Philippine Deaf Research Center and Philippine Federation of the Deaf. (2004). *An Introduction to Filipino Sign Language: (a). Understanding structure*. An Introduction to Filipino Sign Language. Philippine Deaf Resource Center, Incorporated.

Smolyanskiy, N., Huitema, C., Liang, L., and Anderson, S. E. (2014). Real-time 3d face tracking based on active appearance model constrained by depth data. *Image and Vision Computing*, 32(11):860–869, November.

Von Agris, U., Knorr, M., and Kraiss, K.-F. (2008). The significance of facial features for automatic sign language recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE.

Zhang, Z. (2012). Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia*, 19, April.