October 2019

Multimodal Emotion Recognition Using 3D Facial Landmarks, Action Units, and Physiological Data

Diego Fabiano
University of South Florida

Multimodal Emotion Recognition Using 3D Facial Landmarks, Action Units, and Physiological Data

by

Diego Fabiano

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Shaun Canavan, Ph.D.
Tempestt Neal, Ph.D.
Paul Rosen, Ph.D.

Date of Approval:
October 12, 2019

Keywords: Affective, face, detection, expression,
classification, statistical model

Dedication

In this section, I want to dedicate this work to the people that have always supported me. Family has always been there for me; and made it possible for me to move to the United States. Coming from such a small and struggling country like Venezuela, it took all their effort to be able to make this possible. Without their continuous financial and emotional support, I could have never done it.

.

Acknowledgments

In this section, I want to thank the people that helped me along the way of my successful research and educational career. Many colleagues were a fundamental part of my research and education; however, I want to specifically thank David Sherrier for all the years of continuous support and friendship.

Faculty was the main reason of my school success; first, I want to thank Dr. Ali Yalcin for trusting me with giving me my first research opportunity when I barely knew anything. The work I did at home sense was the fundamental building block that allowed me to get an opportunity with Dr. Shaun Canavan.

Lastly, my major professor, Dr. Shaun Canavan has supported me in any way possible; and for that, I will always be extremely grateful. From academic advise, to financial support through funding of my tuition, to trusting me with a Research Assistant position, to nominating me for the Outstanding Undergraduate Research Award, to help with my CPT/OPT applications, and even helping with employment. Dr. Canavan has always gone above and beyond of what is expected of him, thank you.

Table of Contents

List of Tables ii

List of Figures iii

Abstract iv

Chapter 1: Introduction 1
- 1.1 Motivation and Problem Statement 1
- 1.2 Contributions 2

Chapter 2: Related Works 4

Chapter 3: Data Selection and Feature Extraction 7
- 3.1 3D Facial Data 7
- 3.2 Action Units 9
- 3.3 Physiological Data 9

Chapter 4: Experimental Design and Results 11
- 4.1 Feature Analysis 11
- 4.2 Action Units 12
- 4.3 Physiological Data 12
- 4.4 3D Facial Data 13
- 4.5 Emotion Recognition 14
- 4.6 Unimodal vs. Multimodal Emotion Recognition 15
- 4.7 Comparisons to State of The Art 19

Chapter 5: Conclusion and Future Work 20

References 22

List of Tables

Table 1: PCA Rankings 16

Table 2: Unimodal emotion recognition from BP4D+. 17

Table 3: Multimodal emotion recognition from BP4D+. 17

Table 4: Confusion matrix (action units) 18

Table 5: Confusion matrix (physiological data) 18

Table 6: Confusion matrix (action units and physiological) 18

Table 7: Comparison to the state of the art on BP4D+ (recognition accuracy) 19

List of Figures

Figure 1: 3D facial landmarks on corresponding 3D mesh 8

Figure 2: 8 physiological signals used from BP4D+ 13

Figure 3: Top 5 selected 3D facial features across the 4 emotions 14

Abstract

To fully understand the complexities of human emotion, the integration of multiple physical features from different modalities can be advantageous. Considering this, this thesis presents an approach to emotion recognition using handcrafted features that consist of 3D facial data, action units, and physiological data. Each modality independently, as well as the combination of each for recognizing human emotion were analyzed.

This analysis includes the use of principal component analysis to determine which dimensions of the feature vector are most important for emotion recognition. The proposed features are shown to be able to be used to accurately recognize emotion and that the proposed approach outperforms the current state of the art on the BP4D+ dataset, across multiple modalities.

Chapter 1: Introduction

1.1 Motivation and Problem Statement

Affective Computing has been an exciting and growing field in the past two decades and has important applications in artificial intelligence (AI), as being able to recognize emotion is an important part of human intelligence [1]. The ability to recognize emotion has broad impacts for real-world applications in fields as diverse as medicine, defense, entertainment, and retail. Some of these applications include pain recognition [2], customer feedback [3] , and educational video games [4]. To move forward with developing these applications, we need to understand the foundation of autonomy, as well as advance interfaces between human and machines. To do this, we must first understand emotions role in autonomy, including what exactly emotion is. This is a difficult problem as there are currently around 100 definitions of what emotion is [5]. Considering this, there has been a great deal of research into human emotion recognition (HER) in the past decades, where many important advances have been made. This is due in part to the new availability of large, varied, and challenging datasets [6], [7], [8], [9], [10], [11], [12], [13], [14], [15].

This research is fundamental to the current state of Computer Science, specifically, to Machine Learning. Currently, above from the areas of industry previously mentioned, there has being a huge growth in the interest and consideration of mental health as a world community. To fully understand and improve mental health we will need to closely understand emotions, which is a difficult task even for humans. The power of automation has increased exponentially in the past years, and we know have the tools to have machines constantly do incredible calculations.

This way if we get machines to fully understand emotions, we will make great progress towards preventing mental health issues. The research done in this thesis, contributes greatly to the mentioned topic, is a big step and effort in beginning to make machines understand and classify human emotion recognition.

Motivated by this, in this thesis we conduct experiments on multimodal emotion recognition using 3D facial data, physiological signals, and action units from the BP4D+ multimodal emotion corpus [16]. The 3D facial data is comprised of 83 facial landmarks on the face, which examples of can be seen in Figure 1. The physiological signals consist of 8 total signals, per subject, across 4 data types; (1) blood pressure; (2) heart rate; (3) skin conductivity; and (4) respiration. Action units are markers that detail if specific facial muscles have moved [17]. We use this data to conduct both multimodal and unimodal experiments on emotion recognition by training a random forest machine learning classifier [18] to learn four emotions, namely happy, embarrassed, pain, and fear.

1.2 Contributions

The main contribution of this work is an investigation into multimodal emotion recognition using 3D facial data, physiological signals, and action units. The contributions of this thesis are 5-fold and can be summarized as follows.

1. We propose a multimodal approach to emotion recognition using 3D facial data, physiological signals, and action units. To the best of our knowledge, this is the first work to propose a multimodal approach using these modalities.
2. We give a detailed analysis of 3D facial data, physiological signals and action units as they relate to investigating emotion. Each modality, both independently and combined at the feature level (unimodal vs. multimodal) is extensively analyzed.

3. Details on whether 3D facial data, physiological signals, or action units have the greatest impact for positively influencing emotion recognition studies are provided.
4. The efficacy of the approach is tested on the BP4D+ [36] multimodal emotion corpus, outperforming current state of the art approaches across multiple modalities.
5. To the best of our knowledge, this is the first work to report emotion recognition results only using action units from BP4D+.

Chapter 2: Related Works

There is a large and varied body of work into facial expression recognition. Using a Spatio-Temporal Hidden Markov Model (HMM), the intra and inter frame information can be used for this task [19]. It has been shown that using a random forest [18] along with a Deformation Vector Field [20], to obtain the local deformations of the face over time can be used to accurately classify expressions.

Facial expressions have also been successfully classified using a Support Vector Machine (SVM) with a radial basis function (RBF) kernel with geometrical coordinates, as well as the normal of the coordinates [21]. Lucey et al [22], analyzed videos of patients with shoulder injuries to automatically recognize pain. In this work, an Active Appearance Model [23] was used to detect Facial Action Units to distinguish pain on facial expressions. They detail 84.7 for area under the ROC curve on the UNBC-McMaster Shoulder Pain Database [24]. This study is encouraging as it suggests Action Units can be used to recognize emotions (e.g. pain). Deep learning has shown recent success in expression recognition. Using a Boosted Deep Belief Network, Liu et al. [25] trained feature learning, selection, and classifier construction iteratively in a unified loopy framework; which showed an increase in the classification accuracy. Motivated by the Generative Adversarial Model [26], a De-expression Residue Learning [27] approach was proposed which can generate a corresponding neutral expression given an arbitrary facial expression from an image. Yang et al. [28] proposed regenerating expression from input facial images. By using a conditional GAN [29], they developed an identity adaptive feature space that can handle variations

in subjects. Facial expression recognition is a popular approach to recognizing emotion, however, there is also a varied body of work that makes use of multimodal data for emotion recognition.

Soleymani et al. [11] incorporated electroencephalogram, pupillary response, and gaze distance information from 20 videos. They used this data to train an SVM to classify arousal and valence for 24 participants. Kessous also showed an increase of more than 10% when using a multimodal approach [30]. They used a Bayesian classifier, and fused facial expression with speech data that consisted of multiple languages including Greek, French, German, and Italian. Poria et al. [31] performed multiple kernel learning to fuse audio, video, and text modalities. They showed that this approach, along with the integration of a convolutional neural network (CNN) and a recurrent neural network (RNN) [32] can increase the accuracy of recognition. Tzirakis et al. [33] used audio and video data to train an end-to-end multimodal emotion recognition system. Using a CNN to extract audio features, a deep residual network to extract video features, and a long short-term memory (LSTM network) [34], they showed improved results on the RECOLA dataset [35]. Kahou et al. [36] explored multimodal deep learning from audio and video data. The used a CNN to extract visual features, a deep belief network to extract audio features, k-mean "bag-of-mouths" model for visual features around the mouth, and a relational autoencoder for spatio-temporal information. Using this approach, they had the winning submission at the 2013 EmotiW challenge [37]. Liu et al. [38] proposed using facial landmarks along with thermal images to provide regularization on learned features in a deep network. This was done to detect action units in RGB images. While these works are encouraging, the motivation for fusing the selected modalities is not strong, as the correlations are not considered. To learn correlations across modalities, Song et al. [39]

investigated Kernel Canonical Correlation Analysis and Multiview Hidden Conditional Random Fields. Using this approach, they showed encouraging results with audio and video data for recognizing agreement and disagreement in political debates.

Chapter 3: Data Selection and Feature Extraction

The use of 3D facial data (landmarks), action units and physiological data is proposed. The 3 modalities were chosen based on their complementary nature. First, given movement, and the shape of the face changes (3D landmarks), a change in the occurrence of action units [40] can be assumed to eventually happen. The complementary modality is also chosen, physiological data, as facial expressions can be faked. It has been observed that people smile during negative emotional experiences [17]. Considering this, physiological data can complement the other 2 modalities for recognizing emotion. To verify the efficacy of the proposed multimodal approach, a suitably large corpus of emotion data is needed that contains 3D facial data, action units, and physiological data. For these experiments the BP4D+ multimodal spontaneous emotion corpus [16] was chosen. In total, there are over 1.5 million frames of multimodal available in the BP4D+. For this thesis, 192,452 frames of multimodal data from all 140 subjects is used. This subset of data contains 4 target emotions that are happiness, embarrassment, fear, and pain. This subset is used as it is largest set of frames, in BP4D+, that are encoded with action units.

3.1 3D Facial Data

For this thesis 83 3D facial landmarks (same as seen in BP4D+) were used to represent the face. Each of the landmarks were detected using a shape index-based statistical shape model (SI-SSM) [41], that creates shape index-based patches from global and local features of the face. These global and local features are concatenated into one model, which is then used along with a cross-correlation matching technique to match the training data to an input mesh model. Examples of detected 3D facial landmarks can be seen in Figure. 1. For the 3D facial data feature vector, the

coordinates of the 3D tracked facial landmarks were directly used; as they can accurately represent the induced expression that can be seen in the entire 3D model, which contains approximately 30k-50k vertices; where the reduced feature vector contains 249 features (83 3D coordinates). Using this reduced feature space (relative to the entire 3D mesh) allows for lower dimensional data, without sacrificing any recognition accuracy.

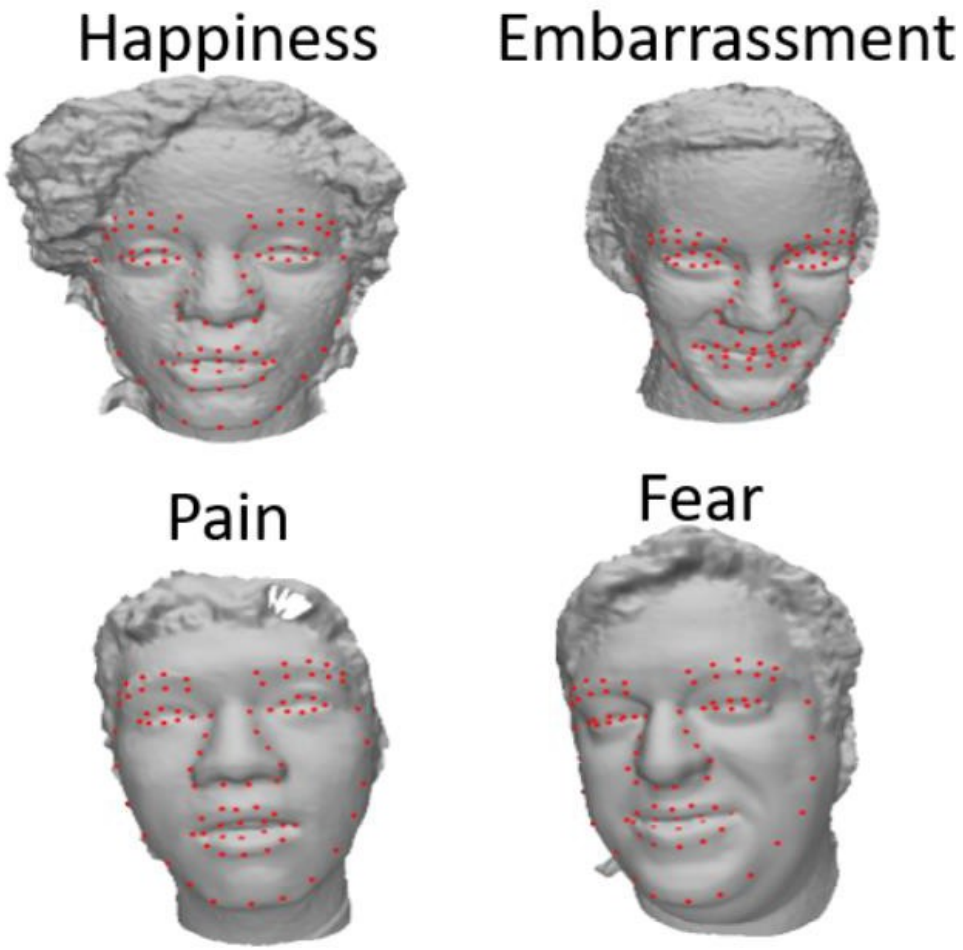

Figure 1. 3D facial landmarks on corresponding 3D mesh. Model for the targeted emotions of happiness, embarrassment, pain, and fear from the BP4D+ [16].

3.2 Action Units

Action units are actions of facial muscles [17], which can be individual or groups of muscles. They can be used to help infer emotion from expression; for example, if action unit 6 + 12 are active, it can be inferred that the subject is happy (i.e. a smile is occurring). For each of the 4 tasks that have action units coding, a total of 35 action units (AUs) were coded by five different expert FACS coders. For each task of all 140 subjects approximately 20 seconds of the most expressive part of the sequence was annotated, giving the 192,452 frames of multimodal data that were for this study. For the AU feature vector, the occurrence of all annotated AUs for each frame are included; where 1 corresponds to the AU being present and 0 corresponds to the AU not being present in the current frame. There are some instances in the BP4D+ where the AU occurrence is listed as 9, which is referred to as unknown. For these experiments, 9 is treated as a 0 (i.e. not present).

3.3 Physiological Data

For each subject and task, the BP4D+ contains 8 separate measurements of physiological data derived from blood pressure (BP), heart rate (HR), respiration (RESP), and skin conductivity (EDA). All physiological data was sampled at 1000 Hz which required to synchronize with the available 3D facial data and corresponding action units to have accurate readings for each frame of data. To synchronize this, the first step is to divide the total number of frames of physiological data by the total number of frames of 3D facial data for that task (average sync value). Then use the average value over the average sync value as the new frame. For example, given a task with 1000 frames of 3D facial data, along with 40,000 frames of diastolic BP we would have 40,000/1,000 = 40, resulting in the taking of the average diastolic BP for every 40 frames. Calculating the mentioned average over all 40,000 frames, results in 1000 frames of diastolic BP

matching to the 1000 frames of corresponding 3D facial data. In this same task, there are 400 frames that include both 3D facial landmarks and AUs (frames labeled with task, subject, and frame number). The corresponding frame number is used to extract that exact index from the calculated diastolic BP averages. This gives the resulting 400 frames of synchronized 3D facial data, physiological data, and action units. For the physiological feature vector, the average value of each frame over all eight of the data types is taken (i.e. fuse the signals).

Chapter 4: Experimental Design and Results

4.1 Feature Analysis

Along with the emotion recognition results, this study is also interested in analyzing which modality and features are most important for our 4 target emotions. To do this, principal component analysis (PCA) was used for feature selection, where 95% of the observed variance is kept [42]. This is useful as PCA captures a rank of the most important features by finding a new set of dimensions (features) in which all of them are orthogonal and ranking them by the variance among them. This is powered by the use of eigen vectors and the corresponding eigen values, which are used to define the variance across the features. The top K (number of features) eigen vectors are selected as the new dimensions and the original dimensions are then transformed into the new ones.

This was done for each of the unimodal feature vectors for all the training data, as well as each individual emotion. This was done to analyze which features are important for emotion recognition in a general sense, and for each targeted emotion resulting in a total of 15 total rankings (3 feature vectors for each: happy, embarrassment, pain, fear, and across 4 target emotions). The features were then ranked based on highest eigenvalue. For each ranking we show the top 5 features for each modality, as this approach to reporting the top features is common in other emotion recognition studies [16]. The top selected features are only used for analysis; all these features are normalized, removing the mean and scaling to the unit variance before analysis through PCA occurs.

4.2 Action Units

The top selected action units included the lips, cheeks, nose, and eye/eyebrow regions. Across each of the target emotions, along with all combined emotions the selected AUs were similar. The difference being their rankings change across different emotion (e.g. AU12 was ranked first for happy, while AU12 was ranked second for embarrassed). Table 1, second column, shows the top 5 selected AUs. As can be seen here the top AUs for Happy are 12, 6, 11, and 7. When considering the Emotion Facial Action Coding System [14], which only looks at emotion-related facial action, Happy, is 6+12. This shows a correlation between the PCA rankings and the action units associated with the emotion. The normalized AU distribution across each target emotion were also calculated. This showed that while each emotion had similar occurring action units, they varied in distribution, which contributes complimentary information to the other modalities. This can explain the increase in accuracy when a multimodal approach is used (Table 3).

4.3 Physiological Data

Most of the top selected features for physiological data were variations on blood pressure (e.g. diastolic and systolic). Pulse rate was also selected as a top feature for each of the target emotions, however, when all emotions were included in the training data, pulse rate was replaced by EDA. This suggests that skin conductivity is important for recognizing multiple emotions. It is interesting to note that for each of the 4 target emotions, not only were the top selected features the same, they were also ranked in the same order. Although each emotion had the same selected physiological data, they all had large variations in the data between them. This variance in data allows for a high level of recognition accuracy (Table 2). Table 1, third column, shows the top 5 selected physiological signals.

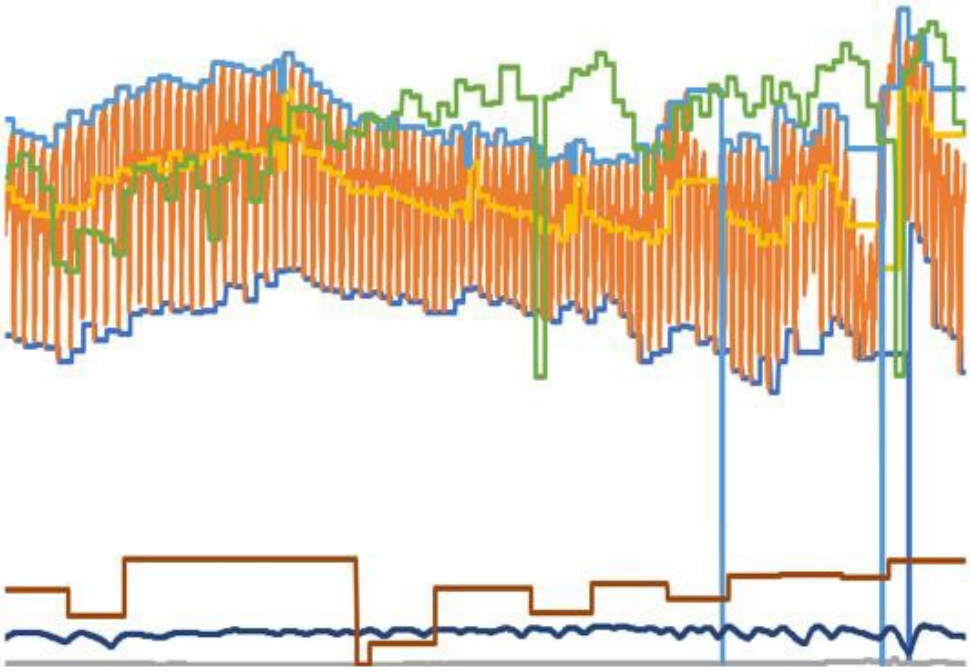

Figure 2. 8 physiological signals used from BP4D+ [48]

4.4 3D Facial Data

When analyzing the 3D facial data, each of the target emotions show variance in the regions of the face that were selected for the top features; which per our analysis makes sense. For example, happy targeted the right eye and eyebrow, and pain was across the right eyebrow, nose, and left eyebrow. These regions of the face are also consistent with the AUs selected as the top features (e.g. mouth, face, eyes/eyebrows). Table 1, last column, details the top 5 selected 3D facial landmarks and Figure. 2 shows an example of the corresponding features (from Table 1) for each of the 4 target emotions, on corresponding 3D mesh models. It can be seen, in Figure. 2, that emotional variance is conveyed in different 3D regions of the face for each of the target emotions. These findings are consistent with our hypothesized relevant features for 3D facial features across the face for emotion recognition.

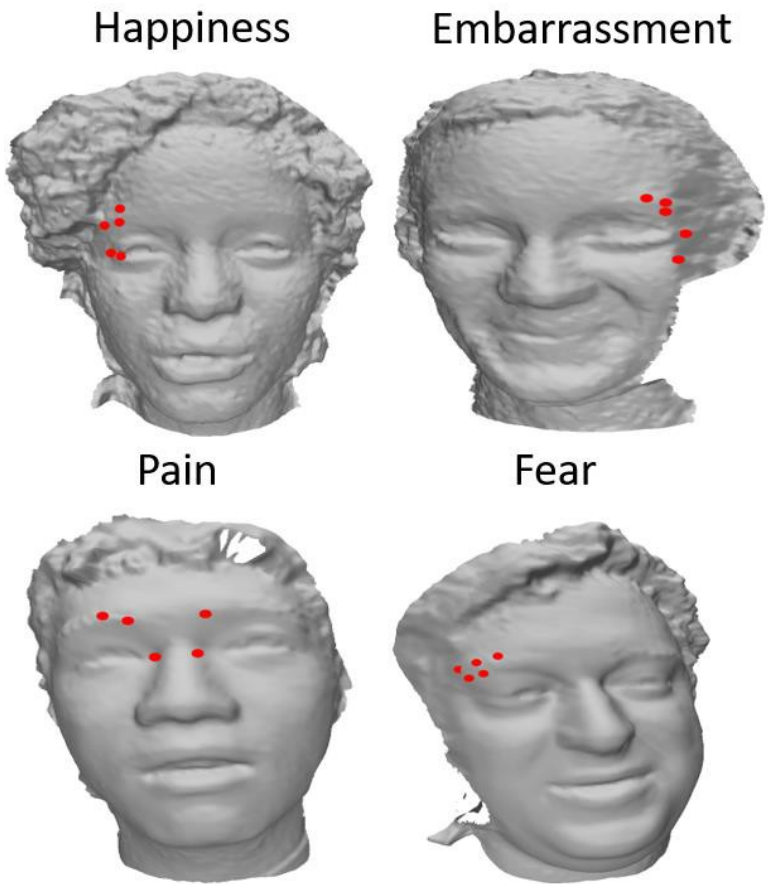

Figure 3. Top 5 selected 3D facial features across the 4 emotions.

4.5 Emotion Recognition

To conduct the emotion recognition experiments, a feature vector for each unimodal and multimodal configuration (Tables 2 and 3) was created. Each of these feature vectors is then used to train a random forest [18] for recognizing the four target emotions. Random forests have successfully been used in a wide variety of classification tasks such as classifying ecological data [43], real-time hand gesture recognition [44], and head pose estimation [45], which makes them a natural fit for the analysis.

4.6 Unimodal vs. Multimodal Emotion Recognition

10-fold cross validation was used for each of the experiments. This type of validation is used commonly, it works by dividing the data in 10 folds or buckets of information equally distributed; after this, the algorithm runs for 10 iterations, in each iteration 1-fold is used for testing and the other 9 for training. This is repeated until each fold has been used for testing once, then the average accuracy of all 10 iterations is taken as the accuracy of the model. This helps to give a more accurate representation of the accuracy of the model, preventing a "lucky split" of the data that would result in unrealistic results.

The results for unimodal and multimodal emotion recognition can be seen in Tables 2 and 3 respectively. When physiological data was used, recognition accuracy was highest for both unimodal and multimodal approaches, achieving an accuracy of 99.94% for the 4 target emotions, with a unimodal approach. This result is intuitive as physiological signals are closely tied to human emotion [46], [8]. For the multimodal feature vectors, when AUs units were combined with physiological data the highest recognition accuracy of 99.95% was achieved. This also agrees with the literature that the fusion of multimodal data, including action units, can provide complimentary information and increase recognition accuracy [47]. Although emotion recognition from AUs shows promising results, especially when fused with other modalities, they exhibit the lowest classification rate of the unimodal feature vectors with a recognition accuracy of 61.94%. The confusion matrices for AUs, physiological data, and AUs combined with physiological data are shown in Tables 4, 5, and 6 respectively (The numbers in each confusion matrix are the total number of frames recognized). Confusion matrices are commonly used to show the accuracy of emotion recognition results. Each row is the ground truth class label (e.g. Happy), each column is the class that the machine learning classifier classified each frame of data as.

\

Table 1. PCA Rankings. For each feature of each individual emotion done along with all 4 target emotions, shown in ranked order.

Emotion	Action Units	Phys	3D Facial Landmarks
Happy	Lip corner puller (12) Cheek raiser (6) Upper lip raiser (10) Nasolabial deepener (11) Lid tightener (7)	Mean BP Diastolic BP Systolic BP Raw BP Pulse Rate	26, 8, 7, 3, 25
Embarrassed	Cheek raiser (6) Lip corner puller (12) Upper lip raiser (10) Lid tightener (7) Nasolabial deepener (11)	Mean BP Diastolic BP Systolic BP Raw BP Pulse Rate	83, 16, 15, 14, 82
Pain	Lip corner puller (12) Cheek raiser (6) Upper lip raiser (10) Nasolabial deepener (11) Lid tightener (7)	Mean BP Diastolic BP Systolic BP Raw BP Pulse Rate	1, 48, 37, 11, 2
Fear	Upper lip raiser (10) Cheek raiser (6) Lid tightener (7) Lip corner puller (12) Nasolabial deepener (11)	Mean BP Diastolic BP Systolic BP Raw BP Pulse Rate	5, 4, 6, 7, 3
All	Lip corner puller (12) Upper lip raiser (10) Cheek raiser (6) Lid tightener (7) Nasolabial deepener (11)	Mean BP Diastolic BP Systolic BP Raw BP EDA	12, 13, 19, 18, 11

Table 2. Unimodal emotion recognition from BP4D+.

	3D Facial Landmarks	**Action Units**	**Physiological**
Accuracy	99.29%	61.94%	**99.94%**
Recall	98.80%	60.35%	**99.95%**
Precision	99.33%	61.00%	**99.95%**

Combining multimodal data has been found to increase emotion recognition including pain in infants [2]. The results in this thesis show similar results with pain as well, increasing from 99.92% with physiological data to 99.98% when AUs were fused with physiological data. It is interesting to note, that while the overall recognition accuracy was higher when AUs were combined with physiological data, the recognition rates for both happy and fear decreased to 99.94% and 99.90% respectively. This can be attributed to some redundant action unit patterns between happy and fear.

Table 3. Multimodal emotion recognition from BP4D+.

	Landmarks + Action Units	**Action Units + Physiological**	**Landmarks + Physiological**	**Landmarks + Action Units + Physiological**
Accuracy	99.53%	**99.95%**	99.76%	99.83%
Recall	99.58%	**99.95%**	99.75%	99.83%
Precision	99.52$	**99.95%**	99.75%	99.85%

Table 4. Confusion matrix (action units).

	Happy	**Embarrassment**	**Fear**	**Pain**
Happy	**32511**	7730	3373	7917
Embarrassment	17561	**26038**	3238	5282
Fear	8773	5206	**14652**	8163
Pain	1983	2334	1685	**46006**

Table 5. Confusion matrix (physiological data).

	Happy	**Embarrassment**	**Fear**	**Pain**
Happy	**51512**	10	5	4
Embarrassment	21	**52080**	4	14
Fear	4	7	**36780**	322
Pain	22	13	6	**51957**

Table 6. Confusion matrix (action units and physiological).

	Happy	**Embarrassment**	**Fear**	**Pain**
Happy	**51504**	10	5	4
Embarrassment	10	**52100**	3	6
Fear	14	16	**36758**	6
Pain	3	9	1	**51995**

4.7 Comparisons to State of The Art

A comparison of the results to the current state of the art was also done. To the best of this study knowledge, these the first experiments to look at combining the modalities, detailed here, from the BP4D+ for emotion recognition. Fabiano et al. [48] proposed a method for emotion recognition using fused physiological signals, and Zhang et al. [16] conducted separate experiments on 3D facial, thermal, and physiological data. Neither group studied the combination of multiple modalities as proposed here. As it can be seen in Table 7, the proposed method outperforms the other methods on each modality that was used (in this thesis), including the overall highest accuracy on the BP4D+. It is important to note the difference in using physiological data compared to Zhang et al [16]. An accuracy of 99.94% was obtained in this thesis, compared to 60.5% with their method. This large difference in accuracy can be attributed to the proposed method, which is the fusion of all 8 signals, from the BP4D+, to train a random forest. Similarly, Zhang et al. used an RBF SVM, however, they used non-fused, handcrafted features compared to the fusion approach that has been presented. The results in this thesis suggest that a fusion-based approach, with physiological data, can lead to an increase of overall emotion recognition accuracy compared to a unimodal approach, which agrees with the work from Fabiano et al. [48].

Table 7. Comparison to the state of the art on BP4D+ (recognition accuracy).

	AU	**3D**	**Phys**	**3D/AU**	**Phys/3D**	**Phys/AU**	**3D/Phys/AU**
Proposed method	61.94%	**99.29%**	**99.94%**	99.53%	99.76	99.95%	99.83%
Fabiano et al. [48]	N/A	N/A	91.59%	N/A	N/A	N/A	N/A
Zhang et al. [16]	N/A	74.8%	60.5%	N/A	N/A	N/A	N/A

Chapter 5: Conclusion and Future Work

A multimodal approach to emotion recognition using 3D facial data, action units and physiological data was proposed. Experiments in both a unimodal and multimodal capacity on four target emotions were conducted. The analysis has shown that 3D facial data shows variations in facial regions allowing for accurate emotion recognition. Physiological data is also shown to be able to be used for emotion recognition due to the changes across emotion. The occurrence of action units shows differences in distribution over 35 AUs across the four-target emotions, which allows for complimentary information to be used when fusing the AUs with other modalities at the feature level. Although the fusion of AUs is shown to increase the accuracy across the four tested emotions, the results also show that directly using AU occurrences without fusing other modalities, for emotion recognition, is still a challenging problem. These results suggest more research is needed to determine the positive impact of using action units in a unimodal approach for emotion recognition.

While these results are encouraging, there are some limitations to the study. First, more multimodal databases need to be investigated, as of this study only made use of BP4D+. Secondly, more details are needed as to why the fusion of AU occurrences showed an increase in accuracy, while using them in a unimodal capacity generated a relatively low accuracy. Lastly, the current study only focused on four emotions. A much larger range of emotions are needed to fully test the efficacy of the proposed approach. Considering this, for the future work, deep neural networks will be used to test the proposed method against different fusion methods including score level fusion, and the fusion of deep and hand-crafted features. A larger set of multimodal datasets will be used,

and the impact of both AU occurrences and intensities for emotion recognition will be investigated. These experiments will be conducted across a larger set of emotions that include, but are not limited to, surprise, sadness, anger, and disgust.

References

[1] R. W. Picard, E. Vyzas and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* 2001.

[2] G. Zamzmi, D. Goldgof, R. Kasturi, Y. Sun and T. Ashmeade, "Machine-based multimodal pain assessment tool for infants: a review," in *arXiv preprint arXiv*, 2016.

[3] E. Cambria, "Affective computing sentiment analysis," *IEEE Intelligent Systems,* pp. 102-107, 2016.

[4] C. Lara, J. Flores, H. Mitre-Hernandwz and H. Perez, "Induction of emotion states in educational video games through a fuzzy control system," *IEEE Transaction on Affective Computing,* 2018.

[5] R. Picard, Affective Computing, MIT Press, 1995.

[6] D. Cosker, E. Krumhuber and A. Hilton, "A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling," *International Conference on Computer Vision,* p. 2296–2303, 2011.

[7] Fanelli et al, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia,* 2010.

[8] Koelstra et al, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing,* 2011.

[9] McKeown et al, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing,* 2011.

[10] A Savran et al, "Bosphorus database for 3d face analysis," 2008.

[11] M. Soleymani, M. Pantic and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE transactions on affective computing,* 2011.

[12] Stratou et al, "Exploring the effect of illumination on automatic expression recognition using the ict-3drfe database," *Image and Vision Computing,* 2012.

[13] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia,* 2010.

[14] L. Yin, Y. Sun, T. Worm and M. Reale, "A high-resolution 3d dynamic facial expression database," *IEEE International Conference on Automatic Face and Gesture Recognition,* 2008.

[15] L. Yin, X. Wei, Y. Sun, J. Wang and M. J. Rosato, "3d facial expression database for facial behavior research," *7th international conference on automatic face and gesture recognition,* 2006.

[16] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Cifci, S. Canavan, M. Reale, H. Horowitz and Yang et al, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[17] P. Ekman, "The argument and evidence about universals in facial expressions," in *Handbook of social psychophysiology*, 1989.

[18] L. Breiman, "Random forests," *Machine learning,* 2001.

[19] Sun et al, "Tracking vertex vlow and model adaptation for 3d spatiotemporal face analysis," *IEEE Transactions on SMC.*

[20] H. Drira, B. Amor, M. Daoudi, A. Srivastava and S. Berretti, "3d dynamic expression recognition based on a novel deformation vector field and random forest," *Proceedings of the 21st International Conference on Pattern Recognition,* 2012..

[21] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah and I. A. Kakadiaris, "3d/4d facial expression analysis: An advanced annotated face model approach," *Image and vision Computing,* 2012.

[22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Mathews.

[23] T. Cootes, G. Edwards and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* 2001.

[24] P. Prkachin and K. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," in *Pain*, 2008.

[25] P. Liu, S. Han, Z. Meng and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets.," in *Advances in neural information processing systems*, 2014.

[27] H. Yang, U. Cliftci and L. Yin, "Facial expression recognition by deexpression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[28] H. Yang, Z. Zhang and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018.

[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *arXiv preprint arXiv*, 2014.

[30] L. Kessous, G. Castellano and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," in *Journal on Multimodal User Interfaces*, 2010.

[31] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," *International Conference on Data MIning,* pp. 439-448, 2016.

[32] S. Bai, J. Kolter and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequnce modeling," *ArVix:1803.01271,* 2018.

[33] P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 8, pp. 1301-1309, 2017.

[34] Hoc, S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation,* pp. 1735-1780, 1997.

[35] F. RIngeval, A. Sonderegger, J. Sauer and D. Lalane, "Introducting the RECOLA mutimodal corpus of remote collaborative and affective interactions," *Face and Gesture Recognition Workshops,* pp. 1-8, 2013.

[36] S. Kahou, P. Bouthillier, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Proumenty, Y. Dauphin, N. Boulanger-Lweandowski and R. Ferrair, "EmoNets: multimodal deep learning approaches for emotion recognition in video," *Journal of Multimodal User Interfaces,* vol. 10, no. 2, pp. 99-111, 2016.

[37] A. Dhall, R. Goecke, J. Joshi, J. Wagner and T. Gedeon, "Emotion recognition in the wild challenge," *ACM International Conference on Multimodal Interaction,* 2013.

[38] P. Liu, Z. Zhang, H. Yang and L. Yin, "Multi-modality empowered network for facial action unit detection," *Winter Conference on Applications of Computer Vision,* 2019.

[39] Y. Song, L. Morency and R. Davis, "Multimodal human behavior analysis: learning correlation and interaction across modalities," *International COnference on Multimodal Interaction,* 2012.

[40] P. Ekman and W. Friesen, "The facial action coding system: A technique for the measurement of facial movement," in *Consulting Psychologists Press*.

[41] S. Canavan, P. Liu, X. Zhang and L. Yin, "Landmark localization on 3d/4d range data using a shape index-based statistical shape model with global and local constraints," *CVIU,,* 2015.

[42] T. Cootes, G. Edwards and C. Taylor, "Active Appearance Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 6, pp. 681-685, 2001.

[43] D. Cutler et al, "Random forests for classification in ecology," in *Ecology*, 2007.

[44] X. Zhao, Z. Song, J. Guo, Y. Zhao and F. Zheng, "Real-time hand gesture detection and recognition by random forest," in *Communications and information processing*, 2012.

[45] G. Fanelli, J. Gall and L. Van, "Real time head pose estimation with random regression forests," in *CVPR*, 2011.

[46] R. B. Knapp, J. Kim and E. Andre, "Physiological signals and their use in augmenting emotion recognition for human–machine interaction," in *Emotion-oriented systems*, 2011.

[47] C. A. Corneanu, M. O. Simon, J. F. Cohn and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence,* 2016.

[48] D. Fabiano and S. Canavan, "Emotion recognition using fused physiological signals," 2019.