Exploring Automatic Emotion Recognition with Neutral Expression Subtraction and a Regression Model

Stef van der Struijk
Department of Artificial Intelligence
University of Groningen, The Netherlands
January 2015
*Supervised by Marco Wiering

***Abstract*—The concept of what an emotion is, is explored by going through the current theories available in the field of emotion studies. These theories are Basic Emotion, Appraisal, Psychological Construction and Social Construction models, which can either be measured in a discrete or scalar manner. Then these theories are used to argue that current classification systems lack the scalability for use in more complex practical situations, as a single label only gives limited information. As answer to this problem, Neutral Expression Subtraction (NES) regression is introduced. This approach uses a neutral expression as baseline and makes it possible to train on any emotion separately by assigning belief values. To test whether this new approach has a similar performance to classification systems, three methods are used on the CK+ database to classify 7 emotions and neutral. Method 1 classifies facial expressions in the classical way. Method 2 classifies facial expressions from data where a neutral expression is subtracted from and Method 3 takes the regression approach and assigns belief values for every emotion to an expression. To compare the methods, Method 3 classifies an expression as the emotion with the highest belief value. The data is transformed by using PCA and either KNN or KNN-regression with K-Means clustering is used. Method 1 achieves 37.7%, Method 2, 45.1% and Method 3, 41.2%. These results suggest that NES is useful for expression recognition and that NES regression, using only data of one emotion, could prove useful for the future. The low overall recognition accuracy is mostly due to the limited use of pre-processing and feature extraction, which should be more advanced in future works to make more conclusive statements about NES regression. NES regression should not be judged by these results alone, because its classification performance was only for comparison reasons. The use of belief values has many more possibilities for complex system behavior than a single classified label.**

I. Introduction

Currently a lot of research is going into developing robots at universities and companies for a variety of purposes. In the future, it's most likely that at some point we will see robots in public areas or even in our homes. To make these robots successful, Human-Robot Interaction (HRI) is of vital importance, if they are going to be a part of our daily lives. An important aspect to make the HRI successful would be the recognition of the emotional state and expressions of the user in such a way that the robot can behave accordingly. This would lessen frustration for the human when interacting with the robot. Also when robots show emotions they are perceived as more friendly and human-like [1]. Humans recognize emotional implicit messages of other humans by observing implicit signals, such as their facial expression, voice pitch and body gestures. Therefore these signals are a good starting point for measuring emotions in humans without the use of intrusive methods. In this paper we focus on facial expressions for emotion recognition, because it has been shown that expressions are reliably associated to subjective experience of emotion and autonomic responses [2]. Before start recognizing emotions or intentions, first a concept of what emotion is and for what purpose we want to recognize emotion, is necessary.

Almost everyone has a basic idea of what an emotion is, however when asked to define emotion, even the literature in the field of emotion studies is divided. There are four main theories: Basic Emotion, Appraisal, Psychological Construction and Social Construction [3]. These theories agree that "*emotion* refers to a collection of psychological states that include subjective experience, expressive behavior (e.g., facial, bodily, verbal), and peripheral physiological responses (e.g., heart rate, respiration)." However besides this agreement, how an emotion manifests itself in a human, differs by which theory is thought to be right. *Basic Emotion* models state that every emotion has a unique mechanism and all "emotions" that don't satisfy a set of criteria are not considered emotions. *Appraisal* models also state that emotions are unique mental states, but don't agree that there are distinct dedicated mental mechanisms. *Psychological Construction* models see mental states as "emerging from an ongoing, continually modified constructive process that involves more basic ingredients that are not specific to emotion". *Social Construction* models view emotions as "social artifacts or culturally-prescribed performances that are constituted by sociocultural factors, and constrained by participant roles as well as by the social context" [3].

Measuring emotion can either be done by discrete categories or scalar dimensions. The use of discrete categories matches most with Basic Emotion theory and for some interpretations of the Appraisal theory. These theories considers 6 emotions to be basic emotions, namely anger, happy, sadness, fear, disgust and surprise (contempt is dubious) [4]. Scalar dimensions are more fit for the Psychological Construction and the Social

Construction theory. These theories view emotions as inter-related entities that differ along global dimensions, such as valence, activity, or approach and withdrawal [5].

The research about whether emotions are universal or vary across cultures, support both views [2]. Many experiments suggest that different cultures are able to understand similar emotional expressions, but there are also cultural differences in the interpretation of an emotional expression and different cultural display rules for showing emotion. Facial expressions are more than just markers of internal states, they show information to others of what the person is planning to do, which is likely to influence the perceiver's behavior by invoking emotion in the perceiver [2].

The aim of this thesis is not to determine the best theory, instead its aim is to find the best possible implementation for real-world systems. Whatever the case, if there are only 6 real basic emotions or that there is a spectrum of emotions, emotions fulfill the purpose of influencing people's own and the perceiver's behavior. Therefore an emotion recognition system should reliably interpret the user's facial expression on what the user's intentions are. Although facial expressions are similar across cultures, there are also cultural and individual differences [2]. Therefore in different cultures, the interpretation of a facial expression is different. This means that there cannot be a single trained system for facial expressions which can identify all emotions of all humans across the globe. Robots should get a grasp on what all these emotions and moods are in certain cultures to successfully respond to humans depending on the context. Therefore, for practical use, the robot should be able to recognize an undefined number of expressions and not only basic emotions, based on the culture that it operates in and act accordingly.

A. *New Approach to Expression Recognition: Neutral Expression Subtraction with Regression*

The process of emotion recognition in facial expressions for automatic systems exists out of three steps: Face Detection, Facial Feature Extraction and Facial Expression Classification. See [6] for an overview. Much of the research has focused on developing novel methods to extract features from either still images or image sequences to improve performance of classification systems. Some methods of still images feature extraction are: Feature-Based [7], and Template-Based such as Active Appearance Models [8] and Gabor wavelets [9]. For Image sequences, some methods of extraction are: Feature Points Tracking [10] and Three-dimensional Models [11]. The last step is the classification of the facial expression. Methods to do so are: Neural-Network-Based, Support Vector Machines and Hidden Markov Models.

Most of these systems are designed with the idea that the ability to classify emotions will somehow improve HRI. However the next step, after classifying an emotion, is how to use this label information to change the behavior of the system, which is often not considered. The human languages contain many labels for affectionate states which by far exceed the 6 basic emotions. To communicate on the same level as humans, the robot should be able to recognize what expression patterns consolidate a certain label, which are not necessary completely isolated from each other. As there are many facial expression patterns a human can make, these labels can contain many of them. Current systems are developed with the databases available, but are not designed with the idea of modification afterwards to include more data or labels. The non-adaptability of these systems to retrain makes it hard to deploy these systems in different settings and cultures. Although some papers have acknowledged this problem and for example designed an adaptive classifier [12]. Also the this-or-that nature of classification does not match the overlapping labels in the human-language. This paper tries to tackle these problems in two steps: Neutral Expression Subtraction (NES) and incorporating NES into a regression model.

1) Neutral Expression Subtraction: NES is the subtraction of a facial expression with a neutral expression. See figure 1 for an example. The idea is that a neutral expression can be used as a baseline to recognize other expressions. Neutral is not considered an emotion, but rather can be seen as a lack of another expression. By subtracting a neutral expression we are left with only what makes that expressions unique, which eliminates the between-subject face-shape differences. Also by taking the maximum difference of an emotional expression and a neutral expression and use that as 100%, it's possible to assign a belief value between 0-100% about how much of an emotion is shown. Subtracting a neutral face has been used for integrating expression-invariant face recognition and identity-independent facial expression recognition [13], but it is unknown to us is if this method has been used in a real-time expression recognition system or for a regression approach.

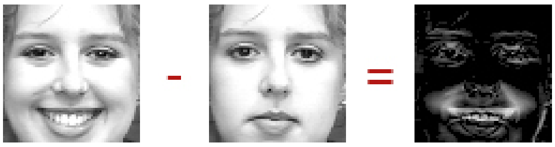

Fig. 1: Neutral Expression Subtraction example

2) NES Regression Approach: By using NES and assigning belief values to expressions, it is possible to take a regression approach to emotion recognition and separate the emotion data per emotion. By doing this, instead of grouping all data together, it is possible to compare an emotion only to a neutral expression. This means that an emotion or expression can be added to a system without influencing the recognition of other emotions or expressions. A new expression is passed through all the previous learned emotions and given per emotion a belief value of how much it resembles that emotion. An ambiguous expression for example can have a high value for two emotions, which is impossible for a classification approach. All these values can be used by a system to respond differently, based on how much it believes that certain emotions are shown, instead of using a single label. This gives a system the potential to be more advanced and adaptive to situations. Also it has the potential to better integrate with other systems when e.g. linked to voice processing. Instead of a label with a weight of how valuable that information is, it can show its certainty by having a high or low belief value. A NES regression approach therefore solves the problem of system adaptability to new data and overlapping human-language labels for expressions.

3) Paper Structure: To test whether a NES regression approach is feasible, three systems are created. Method 1 classifies a full-face in the traditional way, Method 2 classifies in the same way as System 1, but uses NES as input data and Method 3 uses the new NES regression approach. All three methods use K-Nearest Neighbor (KNN) and Method 3 K-Means Clustering as well. This paper first explains how the real-time expression recognition system is built. In section II it is explained how the data is prepared for the three methods. Which is followed by how the three methods work in section III. The results of the three methods are presented in section IV and in section V it ends with a discussion how this method is useful for future use.

II. Neutral Expression Subtraction for Expression Recognition; The System

To explore whether using Neutral Expression Subtraction (NES) with a neutral expression as baseline is a valid option for emotion recognition, a real-time emotion recognition system has been developed in the ROS architecture. This section first explains what database has been used. Then how the images are processed by subtracting a neutral expression and it ends with how the system transforms the data to prepare for machine learning.

A. CK+ database

As database, the Cohn-Kanade AU-coded Facial Expression Database (CK+) is used [14] [15]. The CK+ database contains 593 sequences from 123 subjects. A sequence exists of images ranging from 10 to 60 frames showing the subjects face with the onset of a neutral expression to peak formation of a facial expression. In other words, the frames of a sequence go from a neutral expression to showing a full emotion. Participants age ranges from 18 to 50 years with 69% female, 81% Euro-American, 13% Afro-American, and 6% other groups. For the making of the database, a subject was asked to show an emotion, starting from a neutral expression. Instead of assigning a label based on what the subject is asked to show, the peak frame is reliable FACS coded [16]. The Facial Action Coding System (FACS) classifies emotions based on facial muscles used and is therefore an objective measurement for facial expressions. Some of the sequences showed a dubious emotion according to FACS and therefore those sequences are not labeled as an emotion.

In the system described in this paper, only the sequences accompanied by a FACS determined emotion are used. This totaled to 118 subjects, 327 sequences and 5866 images. The images in these sequences are labeled with a percentage value ranging from 0-100%. The assignment of the percentage is based on the number of images in a sequence, with the first image assigned 0% (neutral expression) and the last image 100%. The images in between are assigned a percentage value in a linear fashion. E.g. the sixth image in a sequence of 11 images is assigned the value 50%. Unfortunately only the peak frame has been FACS encoded and therefore the other images did not contain information about facial muscle contractions and therefore a reliable assignment of a percentage was not possible. Future works could find a more scientific method of assigning a percentage value.

B. ROS node chain

To be able to run a system which recognizes emotions from facial expressions in real time, the Robot Operating System (ROS) architecture, operated in Linux, is used. The images from the CK+ database are loaded into ROS and labeled with the participant number, emotion and percentage value of the shown emotion. These labeled images are converted to gray and face detection is run on it by using OpenCV version 2.9[1] with the package "Haar Cascades". The detected faces are cropped to exclude background and resized to 64x64 pixels. These gray resized faces are saved in a database called *face database* for Method 1, which uses emotion classification in the traditional sense. Then if the face label has a percentage value of 0%, which is considered neutral, the face is temporary stored in memory. Subsequent faces of the same person ranging from 1-100% are subtracted with the neutral face. The gray values of the face images range from 0-255. When the neutral face is subtracted from the face with an expression, the absolute value is taken of the remainder of the image, so that the values again range from 0-255. The remainder of the image is the difference between a neutral expression and an emotion expression, called *difference face* in this paper. The absolute value has been used to be able to show the images and to have more compatibility with OpenCV. Future work could look into how the performance of the system is affected by having negative values for the difference faces. A threshold of 20 is taken for the value of the gray pixels, meaning that every pixel with a value under 20 is set to 0 to reduce noise. These difference faces are saved into a database for training use, called *difference database*.

Only the relevant part of the system used for the results is described here. The complete ROS package is able to detect faces from a video stream (e.g. Kinect, webcam) and can assign a belief percentage for every emotion to a new face by using machine learning on the *difference database*. This ROS package is distributed under the name "Persocom2"[2].

III. Machine learning on the data

For the three methods to either classify an emotion or assigning belief percentages about how much an emotion is shown, Method 1 uses the face database and Methods 2 and 3 use the difference database, described in the previous section. PCA is performed on all the images of both databases, per database. Then the PCA processed data is either directly classified or classification of the highest belief value of all emotions takes place. The results are presented in the next section. A 10-fold cross-validation has been used in which the split is based on the participant number. This means that all images of a particular participant are either in the training

[1] http://opencv.org/

[2] http://ros.org/wiki/persocom

set or in the test set. After describing the methods, this section ends with discussing the special case of when an emotion is labeled as neutral.

A. Data and PCA

For PCA, the python package scikit-learn with the PCA module is used[3]. Every image in a database is first split based on the emotion. For all images, the meta data about what emotion, what percentage of the emotion is shown and the number of the participant is kept. For the data out of the difference database, there has been experimented with adding complete black pictures per person serving as a 0% neutral expression. The idea behind this is that if you would subtract the neutral expression from the neutral expression, you're left with a complete black picture. Then temporary all data is stacked together and transductive PCA with 20 components is performed on all data. The reason that all data is stacked and not that PCA is performed only on the data of a single emotion is because these results were very poor. This calculated PCA matrix is used on all the data for every emotion separately. In the end the system is left with a PCA matrix for new data and PCA processed data for every emotion separately.

B. Classification Method 1 and 2

Method 1 and 2's results are obtained by direct classification using K-Nearest Neighbors as classification algorithm. KNN has been chosen because the output of regression Method 3 should be able to take any value between 0 and 100%. Therefore cluster approaches were out of the question. Also KNN is a simplistic algorithm fitting to this thesis' goal to explore how the regression and classification approach compare. For KNN the python package scikit-learn with the KNeighborsClassifier module[4] is used. As for the parameters, 'weights' has been set to 'uniform', 'algorithm' to 'auto', 'leaf_size' left as default on 30, 'metric' left as default 'minkowski' and 'p' as default 'euclidean_distance'. The influence of the number of neighbors on the percentage correct classified has been explored and is presented in the results section.

The first step is to split the participants in a training group and a test group for every fold in a 10-fold cross-validation. All data points of the training group were saved in a KNN model. Then for all test points the closest k-neighbors were selected based on euclidean distance. Every data point in the model has a classification label. The test data point is classified as the most occurring emotion in the neighbors. Then this classification is validated with what the classification should have been, which is either correct or wrong. The results are presented in the results section as percentage correctly classified.

C. Regression Method 3

Method 3's result are obtained by indirect classification by first obtaining a belief value of every emotion with a regression approach and then classifying it as the emotion with the highest belief value. For the regression approach, the python package scikit-learn with the KNeighborsRegressor module [5] has been used. The same parameters as the classification method have been used, namely 'weights' has been set to 'uniform', 'algorithm' to 'auto', 'leaf_size' left as default on 30, 'metric' left as default 'minkowski', 'p' as default 'euclidean_distance' and the number of neighbors variable.

The images of the participants are split in a training and test set according to a 10-fold cross-validation, based on the participant number. From the training set a KNN-regression model is made for every emotion. Then a test data point is presented to all KNN-regression models. For every model, the k-nearest neighbors are returned with their belief value. The test data point is assigned a belief value based on the average of the k-nearest neighbors. E.g. when there are seven emotions, there are seven KNN-regression models, which gives the test data point a belief value array with a value for every emotion. The idea is when the test expression shows strongly a certain emotion, the KNN-regressor should return neighbors with a high belief value for that emotion, which in turn results in a high belief for that emotion. Other emotions which have different facial expressions should return a low belief value. Then, from all these belief values, the highest percentage is selected and the test data point is classified as that emotion. In this way we can compare the results to the traditional way of classification, while keeping the advantages of regression.

However it turned out that there was a problem that when a brighter image (higher pixel values for the whole image) was presented, the belief values of all emotions were higher. This is not strange considering that higher belief values for an emotion have more white pixels which comes from the differences with a neutral expression. Even though a white image has a big euclidean distance to the data of a certain emotion, the 100% training data points are closer than the 0% training data points. This leads to incorrectly assigning high belief values to test data points far from the trainings data. To solve this problem, K-Means clustering has been used. For K-Means clustering, the python package scikit-learn with the KMeans module[6] with default values and a 'k' value of 1 has been used. After using KNN-regession, the k-nearest neighbors are passed on to the K-Means Clustering, which forms a cluster center of the k-nearest neighbors. Then the average length in euclidean distance from the cluster center to the neighbors is calculated. Also the length from the cluster center to the test data point is calculated. With having both lengths, the idea is that when the length of the test point is multiple times bigger than the average neighbor length, the data point is far removed from that emotion and therefore the value of the belief about the emotion should be lower. For this reason a distance-decay function as seen in equation 1 is introduced.

Equation 1 is a parabolic function with a value between 1 and 0. It starts decaying slowly from 1 and the closer it gets

[3] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[4] http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier

[5] http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html#sklearn.neighbors.KNeighborsRegressor

[6] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans

to r, the faster the function decreases. The -1 is because if the length from the new data point to the cluster center is less than the average neighbor length, the belief value of an emotion should remain unchanged. If the x is smaller than 1 or bigger than r, the multiplier is set to 1 or 0 respectively.

$$\begin{aligned} multiplier &= \sqrt{1-(\tfrac{x-1}{r-1})^2} \\ b &= b * multiplier \end{aligned} \quad (1)$$

where:

$x =$ distance new data point / average distance neighbors
$r =$ radius; how quickly the multiplier decays to 0
$b =$ belief value for an emotion

After the $multiplier$'s changes to the belief values of the emotions, the system checks which emotion has the highest belief value and classifies the new data point as that emotion. Then it checks whether the right emotion is classified or not. The results are presented in the next section.

D. Neutral

Neutral is not part of the basic emotions, therefore in this paper it's considered as lack of another expression. In practical sense this means that everything under a certain belief value of an emotion is considered to be neutral. In this paper the value 20% has been chosen. This 20% is chosen arbitrary, because unfortunately there was not enough time to test the influence of changing this value on the results. This means that for the trainings set of both the classification systems and the regression system that all expressions under a belief value of 20% is relabeled to 'neutral'. For the test set of the classification systems, this means that an expression with mostly neighbors under 20%, is classified as neutral. However for the test set of the regression system, the belief value of all emotions has to be lower than 20% to be classified as neutral.

IV. Results

Table I shows the results of the different methods with different parameters. Method 1 has been tested with different numbers of neighbors. The same is true for Method 2, however this method is also tested whether adding black pictures, which function as a neutral expression with a belief value of 0%, improves the results, which it does. Method 1, which classifies the emotions based on the images from the *face database*, performs better than chance, which is 12.5% with 7 emotions and neutral. Method 2, which classifies the emotions based on the images from the *difference database*, performs better than method 1. Changing the number of neighbors only slightly changes the performance of both methods. The optimal performance of Method 1 is 37.7% and for Method 2 this is 45.1%

Method 3 has been tested by varying the number of neighbors, adding neutral expressions (black images), normalizing the data and using K-Means with different values for r. Normalizing the PCA processed data does not improve the results, but instead decreases the performance by around 10%. Adding neutral expressions (black images) improves the classification results by around 10%. Especially the classification of neutral improves by around 50%. Therefore the other parameters are tested without normalization and with added neutral expressions. Increasing the number of neighbors improves the classification results up until 40 neighbors, after which performance slightly decreases. Performing K-Means after KNN to change the belief value of an emotion, improves the performance by around 15%. This probably means that indeed the problem with KNN with separate emotions was that high belief value data points were selected as the closest neighbor to test data points, even when these test data point were actually far away in Euclidean distance. Varying the r value of the distance-decay function in equation 1, changes the performance by a few percentages. The optimal parameters for Method 3 are 40 neighbors with $r = 3$. The performance is 41.2% classified correctly. In table III the performance with these parameters is shown in a confusion matrix. The table shows that an emotion is often wrongly classified as sad, which could mean that sad is not unique enough. Also the emotions anger, contempt and sad are often wrongly classified as each other. This could mean that these emotions are closely related in expressions. Surprise on the other hand seems to be rather unique, besides being often wrongly classified as neutral.

By comparing Method 1, 2 and 3, we see that the overall performance is not that different. However Method 2 performs the best and Method 3 performs slightly better than Method 1. Another noticeable point between the classification methods and regression method is that the percentage correct classified for the Regression method is more stable than the classification methods. The classification methods have a high rate of classification for certain emotions such as happy and as well for neutral. However contempt, fear and sad are extremely low, which is in contrast with the percentage correctly classified for regression. Something particular between the classification methods and the regression method is that the classification methods score low on contempt, which is not strange regarding the number of examples available, see table II. However regression is scoring rather high for only 179 examples available.

For all systems, the results about the neutral expressions should be interpreted with care, because neutral has a different criteria than the emotions. Also the regression system has a different criteria for an expression being neutral, with all emotions having to be under 20% instead of the most neighbors, therefore the results of the 3 systems are not completely comparable. Take also note of the number of training examples available. The classification system utilizes all training data when classifying. In contrast, regression uses only the training data of 1 emotion.

V. Discussion

This paper started with summarizing that there are 4 main theories about emotions [3]. The Basic Emotion, Appraisal, Psychological Construction and Social Construction theories. Basic Emotion and Appraisal theories fit most the discrete measurement of emotion saying it is either this or that emotion. Psychological Construction and Social Construction theory are usually measured on scalar dimensions, e.g. valence and

	neutral	anger	contempt	disgust	fear	happy	sad	surprise	**average**
1 class noN neig-9	41.9	19.8	23.3	24.7	8.8	54.8	6.1	49.0	**35.5**
1 class noN neig-40	46.2	21.7	27.3	21.0	10.6	58.5	6.7	47.9	**37.7**
2 class noN neig-9 neu	84.3	21.0	11.4	35.8	3.7	68.8	6.2	30.6	**44.1**
2 class noN neig-40 noneu	84.7	21.1	3.3	40.6	10.4	69.7	5.2	23.8	**41.5**
2 class noN neig-40 neu	89.1	25.9	4.2	36.3	4.3	70.9	5.8	24.3	**45.1**
3 reg N neig-40 neu KM-3	31.5	14.9	39.1	58.3	24.9	28.7	16.7	62.6	**30.3**
3 reg noN neig-40 noneu KM-4	3.7	18.4	77.1	45.7	15.3	40.1	25.8	44.9	**30.3**
3 reg noN neig-10 neu KM-4	47.1	18.6	55.9	36.8	17.6	23.6	25.6	40.7	**33.3**
3 reg noN neig-30 neu	52.1	11.3	72.6	26.9	12.2	10.6	16.7	6.4	**23.5**
3 reg noN neig-30 neu KM-4	55.4	19.4	56.4	45.1	18.1	34.2	25.6	43.2	**38.6**
3 reg noN neig-40 neu	27.9	13.2	77.1	30.2	11.3	11.2	19.7	5.3	**25.0**
3 reg noN neig-40 neu KM-2	66.7	13.0	35.2	53.1	16.9	47.3	16.9	23.0	**38.6**
3 reg noN neig-40 neu KM-3	60.9	19.0	52.5	51.3	20.4	42.2	24.2	39.4	**41.2**
3 reg noN neig-40 neu KM-4	57.5	19.8	59.2	48.3	18.1	37.8	25.4	43.2	**40.2**
3 reg noN neig-40 neu KM-5	56.1	19.0	63.7	44.4	16.2	32.1	26.8	42.3	**38.3**
3 reg noN neig-50 neu	49.2	13.8	79.3	32.9	11.7	12.6	19.7	4.3	**24.3**
3 reg noN neig-50 neu KM-4	56.1	19.8	56.4	48.9	16.2	39.8	26.5	42.9	**40.1**

TABLE I: Showing the results of emotion recognition with different parameter values. Every emotion has a different amount of examples, see table II. Therefore the average is a weighted average based on the number of examples.
First column codes: x: Method 1,2 or 3; reg: regression; class: classification; (no)N: (no) normalization; neig-x: number of neighbors; (no)neu: (no) neutral expression added (black images); KM-x: KMeans used with r = x.

	neutral	anger	contempt	disgust	fear	happy	sad	surprise	**Total**
Examples	1308	797	179	669	426	1034	426	1027	**5866**

TABLE II: Number of examples used per emotion. The number of black neutral images has been included for neutral.

		Classified as:							
		neutral	anger	contempt	disgust	fear	happy	sad	surprise
Should be:	neutral	**61**	8	10	4	1	1	14	0
	anger	9	**19**	17	19	4	5	26	2
	contempt	16	9	**53**	0	7	4	12	0
	disgust	8	10	7	**51**	3	5	15	0
	fear	9	24	6	15	**20**	5	17	4
	happy	6	6	8	17	5	**42**	15	1
	sad	10	27	17	10	2	9	**24**	1
	surprise	33	7	2	6	4	4	4	**39**

TABLE III: Confusion matrix of “3 reg noN neig-40 neu KMeans-3” from table I with an average of 41.2% correctly classified.

activity. From a perspective of making a system that can handle complex social interactions with humans, only recognizing 6 basic emotions is too limited. Measuring valence is hard to do objectively, because an expression being negative or positive is a subjective human experience. On the other hand, activity could be measured by muscle activation in a face with e.g. FACS [16]. FACS is usually used for objectively recognizing the basic emotions. As basic emotions are easier to objectively measure, Neutral Expression Subtraction (NES) is in this paper compared with a system recognizing basic emotions.

The task of an expression recognition system is to recognize what makes an expression unique, independent of the person showing the expression. NES eliminates this person dependent difference, because by subtraction the neutral expression of that person, only what makes that expression unique, is left. However the true strength of NES with regression is that it can be extended with any facial expression label without having to re-train the whole system. That is because NES regression makes use of belief values, which are only dependent on facial expressions compared to the person's neutral expression. This eliminates the dependence on the differences between other emotions for recognizing emotions. In the human language, there are many affective labels to describe a person's state of mind, of which some can be recognized in a face. Classification systems are a poor choice if the differences between one label and an other label for an expression are small, as it can only classify it as one of the two. The poor scalability of classification systems limits the performance when adding similar expressions to be recognized. Also most classification systems have to be completely retrained when expressions are added, removed or changed, with the exception of methods such as 1-class classification. This limits the manageability of classification systems in complex environments where adaption can be of vital importance. These limitations are solved by using NES with a regression approach.

To see whether NES is a viable option, three methods have been compared on the CK+ database [14] [15]. Method 1, which uses classification in the traditional sense, obtained 37.7% correctly classified. Method 2, which uses classification on NES data, classified 45.1% correctly. Method 3, which takes a regression approach to the NES data, achieved a 41.2% correct classification. The performance of all three methods are rather poor, but this was to be expected since no feature extraction has taken place. Time limitations made proper feature extraction besides PCA hard to manage, but as all three methods have similar problems, it can still be compared. However these results should be interpreted with care. From these results we can conclude that using NES likely improves results for classification systems. Also regression is not performing much worse, especially when considering that only the data of one emotion is used. This suggests that NES regression has potential for the future, however NES regression has to be tested with proper feature extraction to make conclusive statements. A possible feature extraction method is making a 3D model of the user's face with Active Appearance Models (AAM) [8]. Also using the time component of how the user's expression changes over time could prove to be useful. For classification, only KNN (with K-Means clustering) has been used. The use of other machine learning algorithms could improve results as well. In the current setup, there are many parameters that also can be changed, but have not been due to time constraints or because these parameters are similar for all three systems. These parameters are the resizing of the faces, the number of PCA components, threshold for subtracting faces and absolute value after subtracting a neutral face. A serious limitation is the assignment of belief values to expressions. In this system the belief values are assigned linearly, based on the number of images, however due to time constraints these sequences are not individually checked when the true onset of an emotion takes place. Therefore the belief values are not assigned scientifically credible. Improving these belief values in future works most likely improves the results of the regression method. Another challenge is finding/forming a reliable neutral expression of a person automatically.

The classification results of the regression approach should not be used to judge the use of NES regression, because classification was only done to have a way to compare the methods. In practical situations, the belief values can be used for much more complex behaviors. A system can act with more confidence when the belief value of an emotion is high compared to when it is low. In this way a system can be more precise in its actions depending on what the situation asks for. Also ambiguous expressions can be better interpreted as believing that multiple emotions are expressed than assigning only one label. Using neutral as baseline is not only limited to expression recognition, but could also be used for e.g. physiological signals. The integration with other systems can be more advanced if those systems also have belief values. If for example system 1 classifies some features as angry and system 2 classifies other features as sad, then this information is not very helpful. However if these systems use belief values and system 1 assigns anger as 90% and sad as 60% and system 2 assigns anger as 70% and sad as 72%, then it is more likely that the person is angry, but it still can use the sad information to act more carefully. Therefore by utilizing belief values, made possible by NES, a system has the potential to react in more advanced ways.

References

[1] T. Kishi, T. Kojima, N. Endo, M. Destephe, T. Otani, L. Jamone, P. Kryczka, G. Trovato, K. Hashimoto, S. Cosentino, and A. Takanishi, "Impression survey of the emotion expression humanoid robot with mental model based dynamic emotions," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, May 2013, pp. 1663–1668.

[2] D. Keltner, P. Ekman, G. C. Gonzaga, and J. Beer, "Facial expression of emotion," in *Handbook of Affective Sciences.*, ser. Series in Affective Science, H. H. Goldsmith, K. R. Scherer, and R. J. Davidson, Eds. Oxford University Press, 2003.

[3] J. J. Gross and L. F. Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion review*, vol. 3, no. 1, pp. 8–16, 2011.

[4] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[5] J. A. Russell, "Reading emotion from and into faces: resurrecting a dimensional-contextual perspective," *The psychology of facial expression*, pp. 295–320, 1997.

[6] S.-S. Liu, Y.-T. Tian, and D. Li, "New research advances of facial expression recognition," in *Machine Learning and Cybernetics, 2009 International Conference on*, vol. 2, July 2009, pp. 1150–1155.

[7] M. Pantic and L. Rothkrantz, “Facial action recognition for facial expression analysis from static face images,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 3, pp. 1449–1461, June 2004.

[8] F. Tang and B. Deng, “Facial expression recognition using AAM and local facial features,” in *Natural Computation, 2007. ICNC 2007. Third International Conference on*, vol. 2, Aug 2007, pp. 632–635.

[9] J. Yu and B. Bhanu, “Evolutionary feature synthesis for facial expression recognition,” *Pattern Recognition Letters*, vol. 27, no. 11, pp. 1289 – 1298, 2006, evolutionary Computer Vision and Image Understanding. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865505003478

[10] Y. Zhang and Q. Ji, “Active and dynamic information fusion for facial expression understanding from image sequences,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 5, pp. 699–714, May 2005.

[11] T.-H. Wang and J.-J. J. Lien, “Facial expression recognition system based on rigid and non-rigid motion separation and 3d pose estimation,” *Pattern Recognition*, vol. 42, no. 5, pp. 962 – 977, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320308004160

[12] M. Ishii, “Basic research on facial expression recognition model with adaptive learning capability,” in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, Oct 2011, pp. 3298–3303.

[13] S. Taheri, V. Patel, and R. Chellappa, “Component-based recognition of facesand facial expressions,” *Affective Computing, IEEE Transactions on*, vol. 4, no. 4, pp. 360–371, Oct 2013.

[14] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 46–53.

[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[16] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: Research Nexus*. Salt Lake City, UT, USA: Network Research Information, 2002.