2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Intelligent Facial Action and Emotion Recognition for Humanoid Robots

Li Zhang, Alamgir Hossain
Dept. of Computer Science and Digital Technologies
Northumbria University
Newcastle, UK, NE1 8ST
li.zhang@northumbria.ac.uk

Ming Jiang
Dept. of Computing
University of Leeds
Leeds, UK, LS2 9JT.
jm2000cn@yahoo.com

***Abstract*—This research focuses on the development of a real-time intelligent facial emotion recognition system for a humanoid robot. In our system, Facial Action Coding System is used to guide the automatic analysis of emotional facial behaviours. The work includes both an upper and a lower facial Action Units (AU) analyser. The upper facial analyser is able to recognise six AUs including Inner and Outer Brow Raiser, Upper Lid Raiser etc, while the lower facial analyser is able to detect eleven AUs including Upper Lip Raiser, Lip Corner Puller, Chin Raiser, etc. Both of the upper and lower analysers are implemented using feedforward Neural Networks (NN). The work also further decodes six basic emotions from the recognised AUs. Two types of facial emotion recognisers are implemented, NN-based and multi-class Support Vector Machine (SVM) based. The NN-based facial emotion recogniser with the above recognised AUs as inputs performs robustly and efficiently. The Multi-class SVM with the radial basis function kernel enables the robot to outperform the NN-based emotion recogniser in real-time posed facial emotion detection tasks for diverse testing subjects.**

Keywords—facial action; facial emotion recognition; neural network; support vector machine

I. INTRODUCTION

It is envisaged that humanoid robots will play an increasingly active and engaged role in healthcare and educational settings for isolated elderly, autistic children and average users. However, how to conduct real-time efficient emotion detection from affective facial expressions, gestures and speech in a dynamic daily environment is still a challenging task for humanoid robots. This research aims to incorporate physical cues and anatomical knowledge to guide facial emotion recognition for a humanoid robot as the initial exploration. Psychology study in literature on facial expressions and their associations with emotional and cognitive states is thus explored.

Facial Action Coding System (FACS) for measuring and describing facial behaviours has especially drawn our attention [1]. It associated the momentary appearance changes with the action of muscles from anatomical perspective. The system employed Action Units (AUs) which represent the muscular activities to describe and score facial expressions. These AUs are derived from the study of still images and anatomical texts. Most of the Action Units involve a single muscle. However, there are also cases that two or more AUs are used to represent relatively independent actions of different parts of one particular muscle. The FACS has recovered overall 46 Action Units. It provides a versatile method to describe a wide range of facial behaviours, e.g. facial punctuators in conversation and emotional facial expressions [1].

There was also other research that especially identified the mapping between Action Units and emotional facial expressions, such as Facial Affect Scoring Technique [2]. Moreover, in the Specific Affect Coding System (SPAFF) [3], a set of upper and lower facial Action Units was identified from observational research and used to describe a set of frequently used emotional facial expressions. It has employed seven upper facial Action Units and eight lower facial Action Units for the description of diverse emotional facial behaviours. For example, Cheek Raiser (AU6) and Lip Corner Puller (AU12) were used to describe facial expressions of affection and happiness, while Brow Lowerer (AU4), Upper Lid Raiser (AU5) and Lip Tightner (AU23) were used individually or in combination to indicate anger. Their work provided the mapping of the AU combinations with overall five positive and 12 negative facial expressions and served as a theoretical guide for emotional facial expression recognition.

Furthermore, FACS provides an objective approach which describes the truth of human behaviour and is closely related to signals for emotional facial expressions. It is also capable of describing emotion intensities and compound emotions, and distinguishing fake from real emotional expressions. Therefore, this work is motivated to employ FACS as an intermediate channel which bridges raw motion-based facial representations and emotional facial behaviour recognition. In this work, we summarized the mapping of facial expressions of six basic emotions with six upper and 11 lower facial AUs under the guidance of FACS and Coan and Gottman [3], since these 17 AUs closely relate to the expression of the six basic emotions.

In this research, feedforward neural networks are used to build upper and lower facial action analysers and perform the recognition of the 17 AUs automatically from raw facial expressions extracted by a humanoid robot. The upper facial analyser is able to recognise Action Units including Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Brow Lowerer (AU4), Upper Lid Raiser (AU5), Cheek Raiser (AU6) and Lid Tightener (AU7), while the lower facial Action Units detected

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

are: Nose Wrinkler (AU9), Upper Lip Raiser (AU10), Lip Corner Puller (AU12), Lip Corner Depressor (AU15), Lower Lip Depressor (AU16), Chin Raiser (AU17), Lip Stretcher (AU20), Lip Tightner (AU23), Lip Pressor (AU24), Lips Part (AU25), Jaw Drop and Mouth Stretch (AU26/27).

Subsequently, neural networks and Support Vector Machines are respectively used to detect six emotional facial behaviours with these derived upper and lower AUs as inputs. In this work, we detect the following emotions from facial expressions including happiness, anger, disgust, fear, sadness and surprise. The overall development is integrated with a humanoid robot, NAO, to enable it to recognise emotions from facial expressions during natural human robot interaction.

The paper is organised in the following way. Section II discusses related work. We present the humanoid robot platform and the recognition of AUs and emotional facial expressions in Section III. Evaluations are provided in Section IV. We draw conclusion and discuss future work in Section V.

II. Related Work

According to Kharat and Dudul [4], facial expressions contributed to about 55% effect of an emotional expression in social interactions. Therefore significant progress in facial emotion recognition has been witnessed in cognitive, neuroscience and computational intelligence fields in recent years [5]. As mentioned earlier, psychophysical research identified that facial muscular activities that produce momentary changes in facial appearance can be summarized using Action Units. Well-known six basic emotions have been regarded as universally recognisable because of similar muscle movements used for the expressions of these emotions for people from different culture [1].

Cognitive research also indicated that perception of facial emotions was based on a categorical model [6], which has been intensively employed in the machine learning and computer vision field. Especially since thousands of anatomically possible facial expressions can be described using AUs [5], many computational facial emotion recognition studies focused on AU detection. For example, Bartlett et al. [7] explored a diversity of algorithms for the recognition of 17 Action Units, including AdaBoost and Support Vector Machines. Their system was trained on manually FACS-coded images and obtained high agreement levels with human coders for the recognition of the 17 AUs. Cohn et al. [8] focused on the recognition of AU 1+2 (both inner and outer brow raised), AU 4 (inner brows pulled together and lowered), from varied pose, moderate out-of-plane head motion, and occlusion using discriminant analysis, since the chosen AUs played an important role in describing facial emotion expression and paralinguistic communication. Littlewort et al. [9] employed a fully automated facial action coding system implemented using Support Vector Machines in the problem domain of distinguishing real pain from fake pain. Their system automatically detected patterns of AUs involved in both real and fake pain. It identified that AU4 (Brow Lowerer) was used exaggeratedly for the expressions of fake pain, which was consistent with psychological research findings.

Grafsgaard et al. [10] also employed hidden Markov models (HMM) and AUs to reveal patterns of affective tutorial interactions. Facial expressions were analysed by two certified FACS coders from sample human-human tutorial videos. 16 AUs were selected for manual coding. The identified AUs and dialogue acts recovered from human-human tutorial dialogue were then used as observation sequences to build the most optimal HMM. The model was able to identify five frequently occurred patterns for affective tutorial interactions. There are also several automatic facial expression and gesture labeling systems available, such as FaceSense [11], a computational model for mapping video input to FACS labels and affective-cognitive states, and Acume [12], an open-source tool that analysed naturalistic combinations of dynamic face and head movement across large groups of people.

Facial feature extraction also plays a very important role in automatic emotion recognition systems. Motion-based feature extraction was used in the work of Afzal et al. [13]. They employed a face-tracker to generate 24 point-based face representations from posed and naturalistic facial expressions. The derived facial points were then used to generate both stick-figure models and 3D XFace facial animations of real users. These three representations of emotional facial expressions, i.e. the derived point-light displays, stick-figure models and 3D realistic animations, were used to assess their abilities in conveying emotions. Their experiments indicated that the intermediate-level stick-figure models showing the outline of facial expressions were better encoders of emotions than the other two methods. The study revealed that stick-figure models seemed to focus a lot more on emotionally salient movements and ignore other rendering flaws.

Compared to the above categorical model from the cognitive science perspective, neuroscience research suggested that the perception of facial emotions was best to be described as a continuous model [14]. In this model, each emotion was described using characteristics common to all emotions in a multidimensional space. Although this model showed advantages in explaining emotion intensities compared to the previous categorical model, it was still not easy to use it to describe compound emotions. Therefore, Martinez and Du [15] proposed a new theoretical model for the description of multiple compound emotion categories such as happy or angry surprise. Their model aimed to overcome the difficulty that both of the categorical and continuous models encountered. Their proposed method was to define N distinct continuous spaces and linearly combine these several face spaces to recognise compound emotion categories for facial expressions. This new theoretical model pointed out future directions for building new computational models for compound emotional facial behaviour recognition.

Emotional behaviour recognition and generation have also been developed for social robots to enhance human robot interaction. In the work of Cohen et al. [16], dynamic body postures for several basic emotions were created and validated for a humanoid robot. Schaaff and Schultz [17] developed a system to recognise emotion from electroencephalographic signals using SVMs for humanoid robots and achieved a recognition rate of 47.11% on subject dependent recognition. Ge et al. [18] presented an active vision system, including

robust face detection, tracking, recognition and facial expression analysis, as a comprehensive vision package for robots. Hidden Markov model was used for face recognition and Multi-layer Perceptrons were used to recognise facial emotions from the extracted motion-based representations.

The work presented here is also motivated by Ekman's psychological research of emotional facial expressions. It makes attempts to incorporate psychological emotional knowledge for the descriptions of complicated facial behaviour to advise recognition process. Moreover, the above research of Afzal et al. [13] also showed that point-light displays were less intuitive to human perception of emotions. Therefore, Action Units are used as an objective psychological bridge to link the motion-based representation automatically derived by a humanoid robot with the recognition of emotional facial behaviours. The humanoid robot, NAO, used in this research is thus equipped to detect emotions from real-time posed facial expressions.

III. Emotional Facial Expression Recongition for a Humanoid Robot

This research has been dedicated to a humanoid NAO robot platform although our system can also perform facial emotion analysis and recognition independent of the robot platform. The version of the robot used in this research is the latest NAO NextGen, H25. It has C++ SDKs available to enable researchers to develop advanced intelligent components for robot vision, speech and motion processing. The robot has two built-in cameras with one located on its forehead and the other located at the mouth level. These are 920p cameras and able to run at 30 images/second for (up to) 1280x720 images. NAO is able to move its head by 239° horizontally and by 68° vertically, and its camera can see at 61° horizontally and 47° vertically. Therefore it has a great vision of its environment. The robot platform also provides vision APIs for image processing, movement detection and background darkness checking. In this research, the robot currently focuses on the emotional facial behaviour recognition only from the frontal views of users' posed facial expressions although face tracking and detection capabilities are also provided to allow side facial feature tracking and recognition.

In this research, we employ the robot's C++ SDKs and face detection APIs for the emotional facial expression processing. The overall development is built based on *Naoqi* 1.12.5 C++ Linux 64 version. The NAO cross platform SDKs are also installed so that the compiled program is able to run both on computers and the robot. ALFaceDetection API is employed in this research to enable the robot to provide basic facial feature data, including information about shape of the face, an ID number for the face, the score and name of the recognised face. It also generates 31 2D points for face representation including the contour of the mouth (8 points), nose position (3 points), shape of each eyebrow (3 points) and contour for each eye (7 points). Each point is represented by a pair of x and y coordinates. The robot is able to function well for facial data collection from real-time interaction under normal lab lighting condition (e.g. a clear view of users' faces is captured by NAO's camera without any dark shadow).

In this research, a face detection algorithm is first developed to learn new faces using the learnFace() method and also report the number of detected faces using events, FaceDetected. When a face is detected, the activated FaceDetected event calls the 'callback()' function to make further processing about the collected real-time facial data in order to recognise the associated emotions. The following statement is used to activate the 'callback()' function in the algorithm.

fMemoryProxy.subscribeToEvent("FaceDetected", "OnFaceDetection", "callback")

The real-time facial data are provided by this processing with the support of the above robot vision APIs. The intelligent emotional facial expression recognition system presented here is embedded in this callback() function and is developed to accept the motion-based point-light facial feature representations as inputs. This facial emotion recognition system includes two artificial neural network-based upper and lower facial feature analysers to respectively derive upper and lower facial Action Units from the above point-based facial representation.

As mentioned earlier, the recognised six upper facial Action Units by the upper facial analyser include: Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Brow Lowerer (AU4), Upper Lid Raiser (AU5), etc, while the detected 11 lower facial Action Units by the lower facial analyser include: Upper Lip Raiser (AU10), Lip Corner Puller (AU12), Lip Corner Depressor (AU15), Lips Part or Mouth Stretch (AU25/27) etc. Furthermore, these recognised upper and lower facial Action Units are then used as inputs of neural network and SVM based facial emotion recognisers to detect emotions embedded in the real-time facial expressions.

Overall, at both training and testing stages, the face detection algorithm is first used to best locate users' faces and adjust NAO's cameras. Then the frontal images of users' emotional facial expressions are automatically collected as inputs to this intelligent facial emotion recognition system. It has been tested using posed facial expressions currently with the intention to be further extended to deal with spontaneous facial behaviours in the future. The overall algorithm flow is presented in the following.

Algorithm 1. The Callback Method

Input: The user shows a neutral or emotional facial expression.
Output: Emotion embedded in this input facial expression.

Repeat
1. If no face detected, then the robot says "no face detected" and no further processing.
2. If a face is detected, collect facial data (mouth, nose, eye and eyebrow points) from memory using fMemoryProxy. NAO also greets the user.

2.1 Process upper facial data points for both of the eyes and the eyebrows and send them to the upper facial analyser trained with FACS-coded emotional upper facial expressions [see *A* in Section III for details].

2.2 Output the six recognized upper facial AUs from the processing of 2.1.
2.3 Process lower facial data points for mouth and nose and send them to the lower facial analyser trained with FACS-coded lower facial expressions [see *A* in Section III for details].
2.4 Output the 11 recognized lower facial AUs from the processing of 2.3.
2.5 Send both sets of the recognized upper and lower AUs (obtained respectively from 2.2 and 2.4) to both NN and SVM based facial emotion recognisers respectively trained with emotional facial expressions represented by the 17 selected AUs [see *B* in Section III for details].
2.6 Output the recognized emotion for this input facial expression.

until *Process is killed.*

The system main dataflow and functionalities are provided in Figure 1. The overall system architecture is provided in Figure 2. Although other Action Units also play roles in facial emotion recognition, the above selected AUs have been intensively used in theoretical and psychological research [2, 3] for the description of facial behaviours of six basic emotions. Therefore this research uses the above selected 17 AUs as initial exploration. More facial Action Units and their contribution to emotional facial expressions and recognition will be explored in future work.

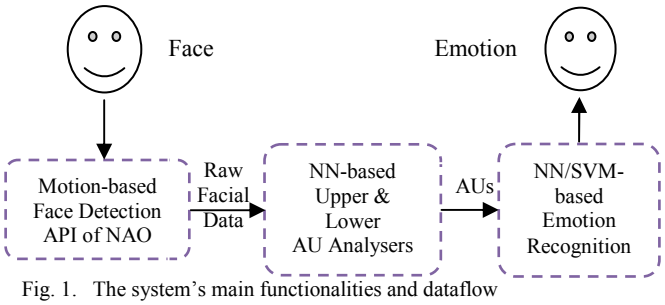

Fig. 1. The system's main functionalities and dataflow

A. Upper and Lower Facial Acton Units Recognition

First of all, diverse emotional facial data are collected from the users. Users are asked to demonstrate the facial expressions of neutral and six basic emotions in front of NAO. Usually the users hold a specific emotional facial expression for about 1-2 seconds. Then the robot will ask users to indicate emotion scores between 0 and 1 for each emotional facial expression under the categories of neutral and the six basic emotions at the training stage.

As discussed earlier, the raw facial data extracted by the NAO robot include a 31-point-based representation for the detailed description of the mouth, nose, both of the eyes and eyebrows. 40 dimensions (20 points) for the description of both of the eyes and both of the eyebrows are used as the input features for the training data set for upper facial Action Unit processing. Then a certified FACS coder provides six scores respectively for the six upper Action Units for the image captured by the robot. Therefore, each training data is represented by a vector of 46 dimensions (40 input dimensions + 6 scores respectively for the 6 upper AUs). 10 emotional facial expressions have been provided by each user and the corresponding scores for the six upper facial AUs are provided by the certified FACS coder. Currently there are 15 users employed to contribute to the training data set construction. There are overall 150 training facial expressions provided by the 15 subjects used for the training of the upper AUs recognition.

In a similar way, mouth and nose data points are automatically extracted from real-time emotional facial expressions. There are 22 dimensions (11 points) used to represent the information of the mouth and nose. The labelling of the selected 11 lower AUs for the training images posed by the users is also provided by the certified FACS coder. Therefore, each training data is represented by a vector with 33 dimensions (22 input dimensions + 11 scores respectively for the 11 lower AUs). Overall 150 training vectors contributed by the same 15 users are used for the training of the neural network-based lower AUs analyser. We also employ 200 images from FACS-coded Cohn-Kanade DFAT-504 (CK+) dataset [19] to further extend the training of the neural network-based upper and lower AUs analysers.

In summary, two feedforward neural networks are developed to learn from the above corresponding training set and to respectively recognise the six upper and 11 lower facial Action Units. Moreover, Backpropagation, as a classic supervised neural network algorithm, is employed in this research. It is chosen due to its promising performances and robustness of the modeling of the problem domain. Moreover, a single hidden layer can approximate any continuous functions. Therefore a model with one single hidden layer is chosen for this application. Both of the upper and lower AUs recognisers are implemented using three-layer neural networks. The three-layer topology of the neural network includes: one input, one hidden and one output layer. Experiments have also been conducted to determine the optimal number of nodes in the hidden layer respectively for the upper and lower AU recognisers. The hidden neuron number for each of the recognisers is determined if the classifier with such a hidden layer is able to produce the most optimal detection results based on the experiments conducted. The following neural network topologies are therefore employed for both of the upper and lower AU recognisers.

The NN for the upper facial AU recognition has 40 nodes in the input layer and six nodes respectively in the hidden and output layers. The 40 nodes in the input layer indicate the facial landmarks for both of the eyes and eyebrows, while the six nodes in the output layer indicate the recognised six upper facial AUs. Similarly, the lower AUs recogniser has 22 nodes in the input layer and 11 nodes respectively in the hidden and output layers. The 22 nodes in the input layer are employed to indicate the data points which represent the mouth and nose positions, while the 11 nodes in the output layer again represent the recognised 11 lower facial AUs.

In order to maintain both of the neural networks generalization capabilities, the training algorithms for both of the upper and lower AUs recognisers minimize the changes made to their corresponding network at each step. This can be achieved by reducing both of the learning rates in the two

training methods. Thus by reducing the changes over time, both training algorithms reduce the possibilities that their corresponding network will become over-trained and too focused on its training set. We adjust parameters such as the learning rate, the momentum and the termination error rate to respectively 0.2, 0.9, and 0.05 for each NN to best achieve a balance between accuracy, speed and generalization performance. After both networks have been trained to reach a reasonable average error rate (less than 0.05), they are used for testing to classify the upper and lower Actions Units from real-time test facial expression data inputs. The employed neural network inference with Backpropagation also proves to have performed efficiently and robustly for AU recognition regardless of the minor changes of the camera angles and scaling differences.

B. NN and SVM based Facial Emotion Recognisers

Cognitive and psychological research has also laid foundations for the mapping between the AUs and emotions embedded in facial expressions [2, 3]. The Specific Affect Coding System (SPAFF) [3] discussed that many AUs can be used individually or in combination to indicate emotional facial behaviours. For example, 'contempt' is closely associated with AU14 (Dimpler), while 'defensiveness' can be physically presented either individually by AU1/AU2 or in combinations of both of them. We have summarised the mapping between the physical cues represented by the 17 selected AUs in this work and the six basic emotions in Table I based on the suggestion of SPAFF. This provides guidance for the recognition of the six basic emotions using the selected upper and lower AUs in this research.

TABLE I. THE MAPPING DERIVED FROM SPAFF BETWEEN THE PHYSICAL CUES REPRESENTED BY SELECTED AUS AND EMOTIONS.

Emotion	Action Units
Happy	**6+12, 12**
Angry	**4+5+7+17+23, 4+5+7+10+23+25**
Sadness	**1+4+15, 6+15**
Disgust	**9, 9+16+15, 9+17, 10+16+25**
Fear	**1+2+4+5+20+25, 1+2+4+5+25**
Surprise	**1+2+5+26/27**

Emotional facial expressions could be diverse and different from one person to another. Table I simply shows some guidance about physical manifest of emotional facial behaviours. The training data used for the training of the upper and lower AUs analysers gathered diverse facial expressions for each emotional category. For example, a facial expression indicates 'happiness' which can be similar to 'affection' (Cheek Raiser + Lip Corner Puller) or very close to a 'positive surprise' (Inner and Outer Brow Raiser, Upper Lid Raiser, Lip Corner Puller and Lips Part).

In order to recognise the six basic emotions from the derived AUs, a supervised neural network facial emotion recogniser is implemented. It also employs a three-layer topology with one input, one hidden and one output layer. Experiments have also been conducted to determine the optimal number of nodes in the hidden layer. Overall, the network accepts the derived 17 AUs as inputs and outputs the recognised six basic emotions. Therefore it has 17 nodes in the input layer and six nodes respectively in the hidden and the output layers. Linear and nonlinear SVM models are also used for the facial emotion recognition in order to identify the most effective classifier for the task.

For the training of the neural network and SVM based facial emotion recognition, the certified FACS coder's annotations of the six upper and 11 lower AUs for the 150 images collected from real subjects are used as the training data. The AU annotations for the selected 200 images provided by the CK+ database are also used for the training of neural network and SVM based emotion detection. Example AU combinations for the emotional facial expressions provided by FACS and SPAFF are also used to complement the training data. Overall the training set contains 380 facial data for emotion recognition.

For the testing of the emotional facial behaviour recognition, another five testing subjects are employed. They are not involved in the training data collection and any algorithm development. For each testing subject, the robot will first of all greet the user and make a brief introduction about what the testing is mainly about. Then the robot requires the user to show a specific emotional facial expression and holds the expression for about one second. Then the real-time system processes the facial data and derives the information for eyebrow points, eye points, mouth and nose for this facial input. Firstly, both the upper and lower facial AU analysers employ the corresponding derived raw facial data points as inputs and output the scores for the selected 17 physical AUs associated with this facial expression. Secondly, these derived values of the 17 AUs are used as inputs for the emotional facial behaviour classifiers. The neural network inference engine and SVM-based models respectively output the detected emotion for this real-time facial expression input. Then the speech synthesis engine of the robot is activated to report the features of the upper and lower face to the testing subject and also inform the user the emotion embedded in his/her real-time input facial expression.

The user then will inform the robot and the researcher who operates the testing if the recognised results are accurate or not based on the user's own interpretation. In the meantime, the robot is in a waiting status until the user is ready to show the next emotional facial expression. Standard emotional facial expressions provided by the FACS are also used to remind the users about any particular emotional expression if help is needed. Otherwise, the users will freely demonstrate their emotional expressions based on their own interpretation during robot human interaction. Figure 2 shows overall system architecture. Overall five testing subjects were involved in the testing and each subject showed five different types of facial behaviours for one specific emotion category. 327 images extracted from CK+ [19] are also used to test the efficiency of facial emotion recognition. Evaluation results for the overall 477 (150+327) facial images are discussed in Section IV.

In order to find the best approach for facial emotion recognition, we also conduct experiments using a multi-class

Support Vector Machine. The algorithm package called $SVM^{multiclass}$ developed by Crammer and Singer [20] is used in this work to recognise emotions from real-time and database input facial data. Similar to the NN-based emotion classifier, it also accepts the 17 selected AUs as inputs and outputs the most likely emotion embedded in the input facial behaviour.

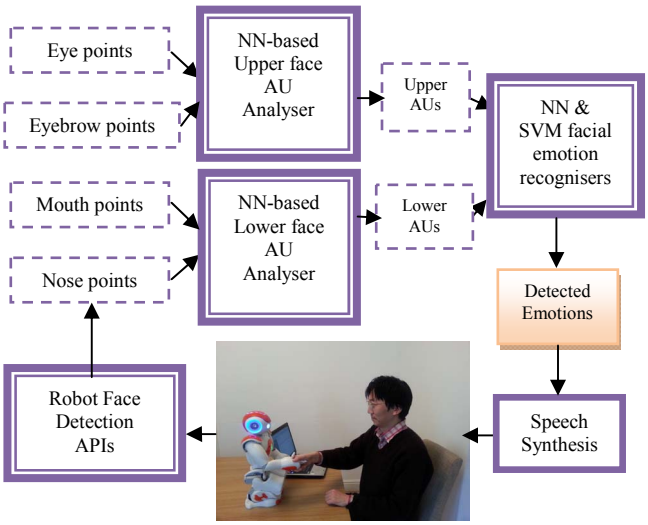

Fig. 2. The system architecture

This multi-class SVM package, $SVM^{multiclass}$, employs a multi-class formulation described in [20]. The latest version $SVM^{multiclass}$ V2.20 has impressive performance. Its runtime increases linearly with the number of training examples. $SVM^{multiclass}$ has a learning module and a classification module. These two modules are employed in this work for the training and testing of a SVM-based emotional facial expression recogniser. This SVM toolkit also provides various parameter options for diverse training and test requirements. For example, kernel function types include linear (default), polynomial, radial basis function and sigmoid. There are also various learning and optimization options including the setting of regularization parameter (i.e. the SVM soft-margin constant, *C*) for the trade-off between training error and margin size (default 0.01). A default setting for the learning method means that a linear kernel and a soft-margin constant of 0.01 are used. In our experiments, in order to find the most effective SVM classifier, different kernel functions, kernels' parameters and soft-margin constants are used. This SVM toolkit is integrated with the robot's C++ platform in Linux in order to perform real-time facial emotion recognition.

Moreover, the previous 380 training examples of the neural network-based emotion classifier are also used to train linear and nonlinear SVMs and the 477 test examples are also employed to test the SVM classifiers.

First of all, four linear kernel based experiments are carried out in order to compare the performance of the four linear classifiers in terms of classification correctness, when a different value of the soft-margin constant is set up for each training session. For a large value of the soft-margin constant, a large penalty is assigned to errors/margin errors while a small soft-margin constant allows constraints to be easily ignored. It is observed that with the substantial increase of the value of this soft-margin constant, *C* (e.g., from 0.01, 1, 500 to 5000), both the number of iterations and number of supported vectors will increase considerably and consequently the corresponding trained classifier is able to improve the accuracy rate significantly. For the evaluation of the test set, we found that the linear SVM trained with the soft-margin constant *C* of 5000 achieved the best performance with a 74.5% accuracy rate.

After we have conducted experiments using the linear kernel with different settings of the soft-margin constants, we move on to employ the polynomial and radial basis function (RBF) kernels whose search space is two-dimensional. For these nonlinear SVMs, the effectiveness of the model not only depends on the soft-margin constant but also relies on the kernels' parameters, such as width of a RBF kernel (*gamma*) or degree of a polynomial kernel (*d*). Both of such kernel parameters decide the flexibility of the resulting classifiers. Experiments have been conducted with the *d* values ranging from one to seven for the polynomial kernel. The degree-2 polynomial with a soft-margin constant, *C*, of 5000 achieves the best performance for the classification of the test data set. It improves the accuracy rate to 77.6%.

Research and experiments also showed that a RBF kernel usually is able to achieve state-of-the-art classification results and outperforms the polynomial kernel [21]. The combination of the soft-margin constant, *C* and the kernel parameter, *gamma*, plays very important roles in affecting the classifier's performance. In order to identify the best pair of these parameters, model selection must be conducted. A grid search on *C* and *gamma* and cross validation have been recommended by state-of-the-art research in the field [22]. Therefore a 5-fold cross validation has been conducted. Such a cross-validation procedure can also avoid overfitting. We use exponentially growing sequences of *C* and *gamma* and employ values ranging from $2^{-5} – 2^{13}$ for *C* and $2^{-5} – 2^{3}$ for *gamma*. The grid search is guided by the cross-validation accuracy (i.e. the percentage of data in each test set correctly classified). The best pair of *C* and *gamma* within the search space in our application is identified as *C*=2000 and *gamma* = 0.25, which yielded the highest cross-validation accuracy. These identified most optimal parameters are then used for the training of the nonlinear SVM model using the overall training data. It has further improved the facial emotion recognition accuracy rate on the overall test set to 81%.

Although such a grid search, especially a complete grid search, for the selection of the best model, sometimes could be time consuming, it provides a safe method for the identification of the best kernel parameter. Since there are only two parameters involved, the computational complexity is not dramatic. There are also other advanced approaches for the search of the best kernel parameters which can reduce computational cost, e.g. by approximating the cross-validation rate. In future work, these advanced methods will be explored. The combination of a coarse grid and a finer grid search on a larger scale of training data will also be employed.

The overall multi-class SVM package is also recompiled under the robot C++ SDK, *Naoqi,* environment and integrated with NAO to perform real-time facial emotion classification. These initial experiments indicated that SVMs show great

flexibility in dealing with complex and challenging facial emotion recognition tasks. Other research [23] also indicated that SVMs are hardly affected by the following two types of errors: "large errors in a small fraction of the data set and small errors in the whole data set". These features provide foundations for a robust and efficient classifier, which worths further exploration.

IV. EVALUATION

As mentioned earlier, five testing subjects were involved in the testing and each subject showed five different types of facial behaviours for each emotion category. Thus 150 test data are employed for the evaluation of the overall facial emotion recognition system during real-time human robot interaction. 327 images extracted from the CK+ database [19] are also used to test the efficiency of facial emotion recognition. The experiments started with natural dialogue based interaction between the robot and the testing subject. It includes greeting, game playing to make users feel relaxed and also allows users to get to know what the testing is about. Then real-time facial expressions are posed by the testing subjects required by the robot. Then the recognised upper and lower facial actions and facial emotions are communicated back to the testing subject by the robot's speech synthesis engine.

The NN-based upper and lower facial AU analysers are tested against the AU annotations provided either by the certified human annotator or the CK+ database. The upper facial AU recognition achieved 80.73% accuracy, while the lower facial AU detection achieved a 77.93% recognition rate. The NN and SVM based facial emotion recognisers have also been tested against the affect labelling results provided either by the testing subjects or CK+. Evaluated using 10-fold cross validation and the test set of 477 images, the neural network-based facial emotion recogniser performed reasonably well with 76% accuracy for the detection of the six emotions.

As discussed earlier, the SVM-based classifiers with linear and non-linear kernels are also employed for the real-time emotion recognition task. The best kernel parameters are also identified for the polynomial and RBF kernels. Especially cross-validation and a grid search are used to identify the most optimal hyperparameters for the RBF kernel. After the determination of the best kernel parameters for the linear and non-linear kernels, 10-fold cross validation has also been conducted using the above test set of 477 images to evaluate SVM-based emotion recognition. As mentioned earlier, the RBF kernel achieves an averaged accuracy of 81%, the best performance, in comparison to the linear (74.5%) and polynomial (77.6%) kernels. The confusion matrix of the RBF based emotion recognition results is also provided in Table II.

The detection results of the neural network and three SVM models all indicate that fear, surprise, sadness and happiness have been well identified followed by 'angry' facial expressions reasonably recognised, while 'disgusting' expressions pose the great challenge to the system. Coan and Gottman's Specific Affect Coding System [3] indicated that disgust facial expressions have physical cues including AU4 (Brow Lowerer), AU10 (Upper Lip Raiser), AU17 (Chin Raiser), AU15 (Lip Corner Depressor) and AU9 (Nose Wrinkler). As mentioned earlier, the testing subjects had the freedom to pose the emotional facial behaviours based on their own interpretation. The images of the users' emotional facial expressions taken during the testing by the robot's Choregraphe software showed that AU10 (Upper Lip Raiser) was used in fearful facial expressions and AU17 (Chin Raiser) was sometimes used in angry facial behaviours. These AUs, i.e. AU10 and AU17, combined with AU4 (Brow Lowerer), make fearful and angry emotional expressions show some resemblance to the disgusting expressions. Therefore, although our system currently is able to detect some key AUs for negative emotional categories, it sometimes has misrecognised disgusting expressions as fearful and angry illustrations.

TABLE II. THE CONFUSION MATRIX FOR THE RBF-BASED EMOTION RECOGNITION

	Surp. (%)	Fear (%)	Anger (%)	Disg. (%)	Sadn. (%)	Happ (%)
Surp.	**83**	10	0	7	0	0
Fear	0	**95**	0	5	0	0
Anger	0	10	**75**	15	0	0
Disg.	0	15	15	**70**	0	0
Sadn.	0	0	9	8	**83**	0
Happ.	15	5	0	0	0	**80**

Moreover, some negative surprise facial expressions also show physical similarities to fearful and disgusting expressions in our testing. They have been misclassified as one another. There is also strong resemblance between happy and positive surprise facial expressions, which also leads to classification errors. These challenging facial emotion recognition tasks also motivate us to employ theoretical research of Martinez and Du [15] to guide finer classifications of these compound negative and positive facial emotions.

We also compare the results of the RBF kernel with those of Wu et al. [24] and Long et al. [25] in Table III. These related applications are selected because of their state-of-the-art performance and usage of the same database for evaluation. The results in Table III indicate that our RBF-based SVM model outperforms the work of [24] and [25]. It especially performs better for the recognition of fearful expressions.

TABLE III. EMOTION RECOGNITION RESULT COMPARISON BETWEEN THE RBF-BASED SVR AND OTHER RELATED WORK

	Anger	**Disg.**	**Fear**	**Happ.**	**Sad.**	**Surp.**	**Average**
Wu et al. [24]	0.83	0.68	0.67	0.88	0.78	0.88	0.786
Long et al. [25]	0.77	0.71	0.69	0.89	0.85	0.89	0.80
Our Work	**0.75**	**0.7**	**0.95**	**0.8**	**0.83**	**0.83**	**0.81**

Furthermore, since more emotional behaviours can be well described using the AUs we focus on in this work including belligerence (AU1 and AU2), defensiveness (AU1 and AU2 with arm-cross or arm on hips gesture), and threats (AU1, AU2 and AU5 probably with body leaning forward), future work will focus on the identification of these facial expressions. Body language such as gesture is another effective channel to

express emotions and feelings. It is also one of the effective indicators for mood, meaning and motive. Therefore, gesture recognition will be explored to combine with emotions derived from facial expressions to draw more reliable conclusions on the perception of non-verbal emotional behaviours. In future work, we also aim to use principal component analysis to select significant AUs from the 17 or more derived AUs for subsequent NN or SVM based emotion detection.

V. CONCLUSION

We have developed intelligent neural network-based upper and lower facial action analysers to recognise respectively six upper and 11 lower facial Action Units. Then another feedforward neural network and linear and nonlinear SVMs are employed to respectively build the emotional facial behaviour recognisers. Six basic emotions are recognised from posed real-time and database facial images. The overall intelligent facial behaviour recognition system is integrated with NAO Linux C++ SDK. The robot is thus able to perform real-time facial emotion recognition for diverse audiences. The most optimal parameters have also been selected for the nonlinear SVM model, RBF, which has outperformed other types of SVM and neural network based facial emotion recognition.

The research could also be further extended to recognise wider categories of AUs and facial emotional behaviours such as interest, pain and stress. Active learning with margin sampling will also be employed in future development to improve the system's performance. Especially it is able to deal classification tasks with limited labelled training data. It employs selective sampling techniques to make attempts to pick the most informative unlabeled examples to enrich the training set. Knowing the true labeling of such ambiguous examples will help the classifier to greatly improve its performance with comparatively less training cost. Thus it is capable of achieving significant performance using fewer unlabelled instances comparing with systems using randomly picked unlabeled samples. It shows great potential in leading to the development of a much efficient robust intelligent facial emotion recogniser to benefit natural human robot interaction.

REFERENCES

[1] P. Ekman, W.V. Friesen, and J.C. Hage, "*Facial Action Coding System*", A Human Face, 2002.

[2] P. Ekman, W.V. Friesen and S.S. Tomkins, "*Facial Affect Scoring Technique: A field study*". Semiotica, 1971, 3(1), 37-58.

[3] J.A. Coan and J.M. Gottman, "The Specific Affect Coding System". In J.A. coan and J.J.B. Allen (Eds.) *Handbook of Emotion Elicitation and Assessment*, pp106-123, New York, NY: Oxford University Press. 2007.

[4] G.U. Kharat and S.V. Dudul, "Human Emotion Recognition System Using Optimally Designed SVM With Different Facial Feature Extraction Techniques". PhD. Anuradha Engineering College; Amravati University, India. (2008).

[5] Z. Zeng, M. Pantic, G.I.Roisman, T.S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol31(1), 2009.

[6] P. Ekman and E.L. Rosenberg, "*What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*". Oxford University Press, New York, 2nd edition, 2005.

[7] M.S. Bartlett, G. Littlewort, M. Frank,C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior". *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '05)*, pp. 568-573, 2005.

[8] J.F. Cohn, L.I. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic Analysis and Recognition of Brow Actions and Head Motion in Spontaneous Facial Behavior". *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC '04)*, vol. 1, pp. 610-616, 2004.

[9] G.C. Littlewort, M.S. Bartlett, and K. Lee, "Faces of Pain: Automated Measurement of Spontaneous Facial Expressions of Genuine and Posed Pain". *Proc. 9th ACM Int'l Conf. Multimodal Interfaces*, 15-21, 2007.

[10] J.F. Grafsgaard, K.E. Boyer, and J.C. Lester, "Toward a Machine Learning Framework for Understanding Affective Tutorial Interaction". In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems*, Chania, Greece, 52-58. 2012.

[11] R. Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures". *Real-time vision for human-computer interaction*, pp. 181–200, 2005.

[12] D. McDuff, R.E. Kaliouby, K. Kassam and R.W. Picard, "Acume: A new visualization tool for understanding facial expression and gesture data". In *Ninth IEEE International Conference on Automatic Face and Gesture Recognition*, USA. pp 591-596, 2011.

[13] S. Afzal, T.M. Sezgin, G. Gao and P. Robinson, "Perception of Emotional Expressions in Different Representations Using Facial Feature Points", In *Proceedings of Affective Computing & Intelligent Interaction*, 2009.

[14] J.A. Russell, "Core affect and the psychological construction of emotion". *Psychological Review*, 110:145–172, 2003.

[15] A. Martinez and S. Du, "A Model of the Perception of Facial Expressions of Emotion by Humans: Research Overview and Perspectives". *Journal of Machine Learning Research*. 1589-1608. 2012.

[16] I. Cohen, R. Looije and M.A. Neerincx, "Child's recognition of emotions in robot's face and body". In *Proceedings of the 6th international conference on Human-robot Interaction*. pp. 123-124. 2011.

[17] K. Schaaff and T. Schultz, "Towards an EEG-based emotion recognizer for humanoid robots". In *Proceedings of IEEE International Symposium on Robot and Human Interactive Communication*, pp. 792-796, 2009.

[18] S.S. Ge, H.A. Samani, Y.H.J. Ong and C.C. Hang, "Active Affective Facial Analysis for Human-Robot Interaction". In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication*, Germany, 2008.

[19] T. Kanade, J.F. Cohn and Y. Tian, "Comprehensive database for facial expression analysis", In *Proceedings of the fourth IEEE International conference on automatic face and gesture recognition*, France, 2000, pp. 46–53.

[20] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multi-class SVMs", *JMLR*, 2001.

[21] A. Ben-Hur and J. Weston, "A User's Guide to Support Vector Machines". *Methods in Molecular Biology*, 609, 223-239 (2010).

[22] C.W. Hsu, C.C. Chang and C.J. Lin, "A practical guide to support vector classification". Updated version. Tech. rep., Department of Computer Science, National Taiwan University. 2010.

[23] R. Hable and A. Christmann, "On qualitative robustness of support vector machines". *Journal of Multivariate Analysis*, 102:993-1007, 2011.

[24] T. Wu, M. Bartlett, and J. R. Movellan, "Facial expression recognition using gabor motion energy filters". In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, IEEE Computer Society Conference on, pages 42–47, 2010.

[25] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett, and G. Littlewort, "Learning spatiotemporal features by using independent component analysis with application to facial expression recognition". *Neurocomputing*, 93:126 – 132, 2012.