Expert Systems with Applications 40 (2013) 5160–5168

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot

Li Zhang [a,*], Ming Jiang [b], Dewan Farid [c], M.A. Hossain [a]

[a] Computational Intelligence Research Group, Dept. of Computing Science and Digital Technologies, Faculty of Engineering and Environment, University of Northumbria, Newcastle NE1 8ST, UK
[b] School of Computing, University of Leeds, Leeds LS2 9JT, UK
[c] Dept. of Computer Science & Engineering, United International University, Bangladesh

ARTICLE INFO

Keywords:
Facial emotion detection
Action Units
Latent Semantic Analysis
Human robot interaction

ABSTRACT

Automatic perception of human affective behaviour from facial expressions and recognition of intentions and social goals from dialogue contexts would greatly enhance natural human robot interaction. This research concentrates on intelligent neural network based facial emotion recognition and Latent Semantic Analysis based topic detection for a humanoid robot. The work has first of all incorporated Facial Action Coding System describing physical cues and anatomical knowledge of facial behaviour for the detection of neutral and six basic emotions from real-time posed facial expressions. Feedforward neural networks (NN) are used to respectively implement both upper and lower facial Action Units (AU) analysers to recognise six upper and 11 lower facial actions including Inner and Outer Brow Raiser, Lid Tightener, Lip Corner Puller, Upper Lip Raiser, Nose Wrinkler, Mouth Stretch etc. An artificial neural network based facial emotion recogniser is subsequently used to accept the derived 17 Action Units as inputs to decode neutral and six basic emotions from facial expressions. Moreover, in order to advise the robot to make appropriate responses based on the detected affective facial behaviours, Latent Semantic Analysis is used to focus on underlying semantic structures of the data and go beyond linguistic restrictions to identify topics embedded in the users' conversations. The overall development is integrated with a modern humanoid robot platform under its Linux C++ SDKs. The work presented here shows great potential in developing personalised intelligent agents/robots with emotion and social intelligence.

1. Introduction

It is a long-term research goal for human computer interaction to build human-like computer interfaces. This endeavour has given rise to agent-based user interfaces (Endrass et al., 2011; Zhang et al., 2009; Zhang and Barnden, 2012). Moreover, we believe it will make intelligent agents possess human-like behaviour and narrow the communicative gap between machines and human-beings if they are equipped to interpret human emotions during social interaction. Thus in this research, we equip our AI agent with emotion and social intelligence. According to Kappas (2010), human emotions are psychological constructs with notoriously noisy, murky, and fuzzy boundaries that are compounded with contextual influences in experience and expression and individual differences. These natural features of emotion also make it difficult for a single modal recognition, such as via acoustic-prosodic features of speech or facial expressions. Since human being's reasoning process takes related context into consideration, in our research, we intend to make our agent take multi-channels of subtle emotional expressions embedded in social interaction contexts into consideration to draw reliable affect interpretation. The research presented here focuses on the production of an intelligent agent with the abilities of inferring emotions from facial expressions and interpreting dialogue contexts semantically to stimulate human robot interaction as our initial exploration.

Motivated by cognitive and psychological research (Ekman and Friesen, 1976), this work first aims to incorporate physical cues and anatomical knowledge of facial behaviour to guide emotional facial expressions recognition. In order to achieve this goal, first of all, psychology study in literature on facial expressions and their associations with emotional and cognitive states is explored. Facial Action Coding System (FACS) for measuring and describing facial behaviours has especially drawn our attention (Ekman and Friesen, 1976). It associated the momentary appearance changes with the action of muscles from anatomical perspective. FACS employed Action Units which represent the muscular activities to describe and score facial expressions. Most of the Action Units involve a

* Corresponding author. Tel.: +44 191 243 7089.
E-mail addresses: li.zhang@northumbria.ac.uk (L. Zhang), m.jiang@leeds.ac.uk (M. Jiang), d.farid@northumbria.ac.uk (D. Farid), alamgir.hossain@northumbria.ac.uk (M.A. Hossain).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.03.016

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. 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 (Ekman et al., 2002).

Related research which identified the mapping between Action Units and emotional facial behaviours was also well established, such as Facial Affect Scoring Technique (Ekman et al., 1971) and Facial Action Coding System Affect Interpretation Dictionary (Ekman et al., 2013). In the Specific Affect Coding System (SPAFF) (Coan and Gottman, 2007), a set of upper and lower facial Action Units was used to describe several frequently used emotional facial expressions. It employed seven upper and eight lower facial Action Units for the description of five positive and 12 negative facial expressions. For example, Cheek Raiser (AU6) and Lip Corner Puller (AU12) were used to describe facial expressions of affection and happiness. Their work also served as a theoretical guide to our research.

As discussed above, FACS provides an objective approach which describes the truth of human behaviour and is closely related to physical indicators of emotional facial expressions. It is also capable of describing emotion intensities and compound emotions, and distinguishing fake from real emotional expressions. Therefore, this work employs FACS as an intermediate channel to link raw motion-based facial representations to emotional facial behaviour recognition.

In this work, supervised neural networks are used to build upper and lower facial action analysers, which recognise 17 Action Units automatically from raw facial data collected by the robot vision APIs. A neural network-based facial emotion recogniser is also implemented to identify seven basic emotions from facial behaviours with the 17 derived upper and lower AUs as input features. The detected emotions include neutral, happiness, anger, disgust, fear, sadness and surprise. The overall development is applied to a humanoid robot platform.

Moreover, in order to enable the robot to perform natural dialogue-based interaction with users based on the affective facial expression interpretation, Latent Semantic Analysis (LSA) is also used to calculate semantic similarities between users' utterances and the training corpus with clear conversational themes to identify conversational themes embedded in the users' dialogue. The recognised affective facial expressions and detected discussion themes embedded in the dialogue are then used to advise the robot to make appropriate responses to stimulate the interaction. Responding regimes and appraisal rules are also developed for the response generation. The semantic-based topic detection also shows robustness and flexibility in dealing with open-ended topic detection tasks without constraints of scenarios. The overall system architecture is presented in Fig. 1.

Overall, the paper is organised in the following way. Section 2 presents related work. We present the humanoid robot platform and the recognition of 17 AUs and emotional facial expressions respectively in Sections 3 and 4. The semantic-based topic theme detection is presented in Section 5. Evaluation and related discussions are provided in Section 6. We draw conclusion and discuss future work in Section 7.

2. Related work

Significant progress in facial emotion recognition has been witnessed in cognitive, neuroscience and computational intelligence fields (Zeng et al., 2009; Wong and Cho, 2009; Ammar et al., 2010 and Rao et al., 2011). Kharat and Dudul (2008) also claimed that facial expression contributed to about 55% effect of overall

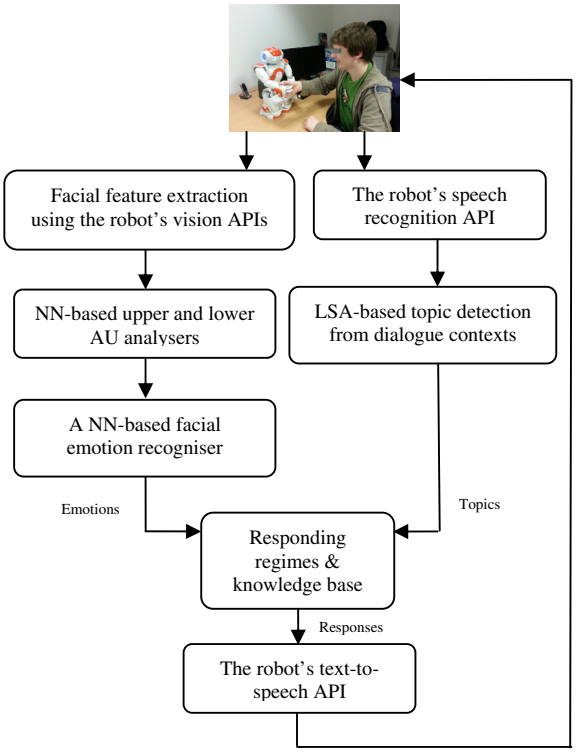

Fig. 1. The overall system architecture.

emotion expression during social interactions. As mentioned earlier, psychophysical research identified that facial muscular activities that produce momentary changes in facial appearance can be summarised using Action Units. Well-known six basic emotions have been regarded as universally recognisable because of similar muscle movements used for the expression of these emotions for people from different culture (Ekman and Friesen, 1976). This research therefore focuses on the recognition of neutral and these six basic emotions with the incorporation of the knowledge of AUs for a humanoid robot.

Perception of facial emotions was regarded to be based on a categorical model in cognitive research (Ekman and Rosenberg, 2005), which has been intensively employed in the machine learning field. Especially since thousands of anatomically possible facial expressions can be described using AUs (Zeng et al., 2009), many computational facial emotion recognition studies employed AUs. For example, Bartlett et al. (2005) explored a diversity of algorithms for the recognition of 17 Action Units, including AdaBoost and support vector machines. Their system was trained with manually FACS-coded images and obtained high agreement levels with human coders for the recognition of the 17 AUs. However, their system did not explore the recognition of emotions from these 17 AUs as the next step processing. Cohn et al. (2004) 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 emotion expression and paralinguistic communication. Littlewort et al. (2007) 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. There are also several automatic facial expression and gesture labeling systems available, such as

FaceSense (Kaliouby and Robinson, 2005), a computational model for mapping video input to FACS labels and affective-cognitive states, and Acume (McDuff et al., 2011), an open-source tool that analysed naturalistic combinations of dynamic face and head movement across large groups of people.

Feature extraction also plays a very important role in automatic facial emotion recognition systems. Motion-based feature extraction was used in the work of Afzal et al. (2009). 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.

Moreover, neuroscience research suggested that the perception of facial emotions was best to be described as a continuous model, compared to the above categorical model from the cognitive science perspective (Russell, 2003). In this model, each emotion was described using characteristics common to all emotions in a multi-dimensional space. Although this model showed advantages in explaining emotion intensities compared to the previous model, it was still not easy to use it to describe compound emotions. Therefore, Martinez and Du (2012) 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.

In order to enhance human robot interaction, emotional behaviour recognition and generation have also been developed for social robots. In the work of Cohen et al. (2011), dynamic body postures for several basic emotions were created and validated for a humanoid robot. Schaaff and Schultz (2009) developed a robotic system to recognise emotion from electroencephalographic signals using support vector machines and achieved a recognition rate of 47.11% on subject dependent recognition. Ge et al. (2008) 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.

As discussed above, 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. (2009) 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 used in this research is equipped to detect emotions from real-time posed facial expressions.

Significant progress in emotion recognition from text and dialogue has also been witnessed by the last decade (Endrass et al., 2011; Ptaszynski et al., 2009 and Zhang et al., 2008). For example, Endrass et al. (2011) carried out study on the culture-related differences in the domain of small talk behaviour. Their agents were equipped with the capabilities of generating culture specific dialogues. There is much other work in a similar vein. Ptaszynski et al. (2009) employed context-sensitive affect detection with the integration of a web-mining technique to detect affect from users' input and verify the contextual appropriateness of the detected emotions. However, their system targeted interaction only between an AI agent and one human user in non-role-playing situations, which greatly reduced the complexity of the modelling of the interaction context. Moreover metaphorical language has been used in literature to convey emotions, which also inspires cognitive semanticists. However, the detection of such metaphorical phenomena posed great challenges to automated linguistic processing tools. In order to identify a few types of metaphorical expressions and go beyond the restrictions of linguistic features, this work is also motivated to employ Latent Semantic Analysis to deal with open-ended topic theme detection.

3. The vision APIs of NAO

This research has employed a humanoid NAO robot platform. The version of the robot used in this research is 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 920 p cameras and able to run at 30 images/s for (up to) 1280×720 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.

The robot's C++ SDKs and face detection APIs are employed for the intelligent facial emotion recognition in this research. 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 a 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 very well for facial data collection from real-time interaction under normal lab lighting condition.

In this research, the face detection algorithm is first developed to learn new faces via 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 of the collected real-time facial data in order to recognise the associated emotions. The 'callback()' function in the algorithm can be activated using the method of MemoryProxy.subscribeToEvent.

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.

We employ the following upper facial Action Units in this research: 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 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), and Lips Part (AU25), Jaw Drop and Mouth Stretch (AU26/27). Then the recognised upper and lower facial Action Units are used as inputs for the neural network based facial emotion recogniser to detect emotions embedded in the real-time facial expressions.

In this research, we use the face detection algorithm first to best locate users' faces and adjust NAO's cameras. Then the frontal views of users' emotional facial expressions are collected by NAO's vision APIs. The overall affective facial behaviour recognition system is then activated to recognise emotions from facial expressions. It is currently designed to recognise posed facial expressions 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: The Callback Method

Input: The user shows an emotional facial expression.
Output: Emotion embedded in this 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 both of the eyebrows and send them to a feedforward neural network classifier trained with FACS-coded emotional upper facial data.
   2.2 Output recognised upper facial AUs from the processing of 2.1.
   2.3 Process lower facial data points for mouth and nose and send them to another feedforward neural network trained with FACS-coded lower facial expressions.
   2.4 Output recognised lower facial AUs from the processing of 2.3.
   2.5 Send the recognised upper and lower AUs (obtained respectively from 2.2 and 2.4) to a neural network based facial emotion recogniser trained with emotional facial expressions represented by the 17 selected AUs.
   2.6 Output the recognised emotion for this facial input data from the processing of 2.5.

until *Process is killed.*

Fig. 2 shows the vision processing functions and dataflow of the emotion recognition process. Although other Action Units also play roles in emotional facial expression recognition, the above selected 17 AUs have been intensively used in theoretical and psychological research (Ekman et al., 1971; Coan and Gottman, 2007) for the description of facial behaviours of six basic emotions. Therefore this research uses these selected AUs as initial exploration. More Action Units and their contribution to facial emotion recognition will be explored in future work.

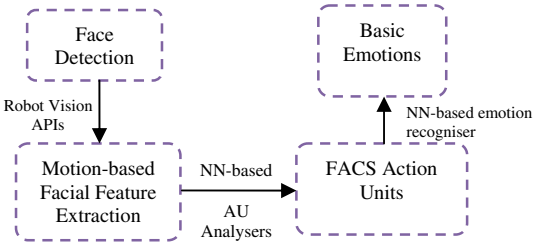

Fig. 2. The main functionalities and dataflow for facial emotion recognition.

4. Facial emotion recognition

For the training of the intelligent facial emotion recognition, emotional facial data are collected from the subjects as the first step. Subjects are instructed to demonstrate the facial expressions of neutral and six basic emotions in front of NAO's cameras. Usually the subjects hold a specific emotional facial expression for about 1–2 s. Then the robot will ask the subjects to indicate emotion scores between 0 and 1 for each emotional facial expression under the categories of neutral and the six basic emotions.

In order to include more diversity and robustness to the training set, a FACS-coded Cohn-Kanade DFAT-504 Facial Expression Database (Kanade et al., 2000; Lucey et al., 2010) is employed. This database contains approximately 2000 sequences of emotional facial images collected from over 200 subjects. We employ the peak-frame facial image of each emotional type from each selected subject to enrich the training set. These selected peak-frame AU-coded images are individually displayed on a desktop screen in front of the robot. NAO is also able to capture each of these 2D database images in real time and generate a 31-point facial feature representation as if the facial expression is posed by a real subject. Fig. 3 shows facial feature capturing using both a real subject and 2D FACS-coded images from the database.

The robot is able to extract raw facial data from both the real-time posed affective facial expressions and the displayed images provided by the Cohn–Kanade database. The extracted facial data by the robot includes a 31-point-based representation for the information of the mouth, nose, both of the eyes and both of the eyebrows as mentioned above. One hundred peak-frame affective facial images contributed by 20 users from the Cohn–Kanade database are used for the training data construction. The image database also provides each peak-frame image with an FACS-coded file. It contains a set of specific AUs and their corresponding intensities labelled by certified FACS coders for each peak-frame emotional facial image. An emotion label is also attached with the peak-frame image for each selected image sequence. These AUs provided by the database and their intensities for the selected 100 emotional images are then used for the training of both upper and lower AU analysers. There are also images which only have identified AUs attached but without any intensity scores available. If this is the case, we then manually provide the value of 0.1 as the intensity score for each identified AU.

For the processing of those real-time posed facial behaviours by the real subjects, a certified FACS coder provides intensity scores respectively for the six upper AUs and 11 lower AUs for each captured facial expression. Five emotional facial expressions have been provided for each emotional category by each user and the corresponding intensity scores for the 17 AUs are provided by the certified FACS coder. Currently there are two users employed to contribute to the training data set construction. Therefore there are overall 70 training vectors collected from real-time posed facial emotional expressions by real subjects. Integrated with the above 100 training images extracted from the image database, we build

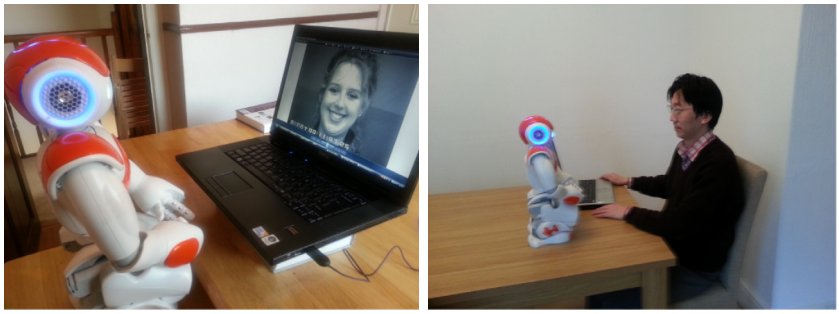

Fig. 3. Training data collection from posed emotional facial expressions of a real subject and the AU-coded database images (©Jeffrey Cohn).

a training set with 170 FACS-coded samples for the training of the upper and lower AUs recognisers.

As mentioned earlier, neural networks are used to implement both upper and lower AUs analysers. They are used to respectively automatically identify the six upper and 11 lower facial AUs in this research. For the training of the upper AU analyser, 40 dimensions (20 points) for both of the eyes and both of the eyebrows are extracted from the 31-point facial representation and used as the input features. The expected intensity outputs of the six upper AUs are taken from the above 170 FACS-coded training samples. Therefore, each training data is represented by a vector of 46 dimensions (40 input feature point dimensions + 6 intensity scores respectively for the 6 upper AUs). In future work, other FACS-coded images and video databases, such as MMI database (Pantic et al., 2005), will be explored to further extend training of the neural network-based upper and lower AUs analysers. The neural network topology of the upper AU analyser is provided in Fig. 4.

Furthermore, mouth and nose data points are also automatically extracted from the training emotional facial expressions. There are 22 dimensions (11 points) used to represent the information of the mouth and nose. The intensity labelling of the selected 11 lower AUs is also provided by the above 170 FACS-coded training data set. Therefore, each training data record is represented by a vector with 33 dimensions (22 input dimensions + 11 intensity scores respectively for the 11 lower AUs). The neural network topology of the lower facial AU analyser is also presented in Fig. 5. In summary, the above two feedforward neural networks are developed to learn from the FACS-coded emotional facial images and to respectively recognise the selected six upper and 11 lower facial Action Units.

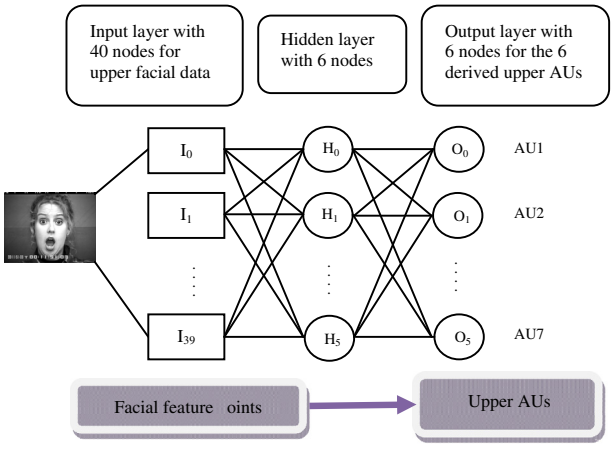

Fig. 4. The topology of the upper facial AU analyser (©Jeffrey Cohn).

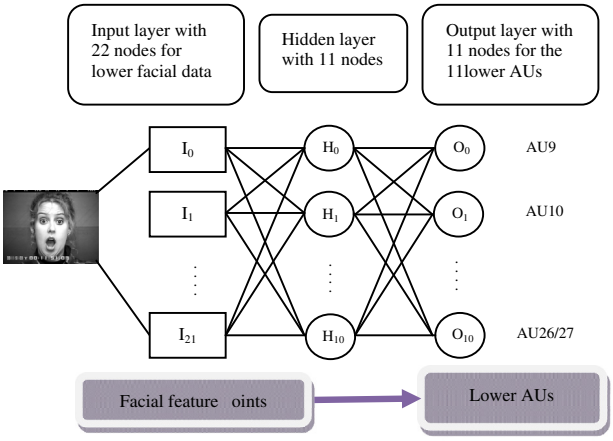

Fig. 5. The topology of the lower facial AU analyser (©Jeffrey Cohn).

In our experiment, we have also made other attempts for the recognition of the some of the AUs. For example, fuzzy logic has also been developed to recognise the openness of the mouth (i.e. AU26/27) and the inner and outer eyebrow raiser (i.e. AU1 and AU2). However, the experiments conducted indicated that the boundary values selected for the fuzzy membership calculations were very sensitive to the distance between the robot and the subjects. However, the feedforward neural network inference with Backpropagation proves to have performed more efficiently and robustly for the AUs recognition regardless of the minor changes of the distances between the robot and the users and the camera angles.

Generally, neural networks are widely used in face recognition research (Zeng et al., 2009). 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 modelling 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 network topologies as shown in Figs. 4 and 5, which include one input, one hidden and one output layer.

As shown in Fig. 3, the neural network for the upper facial AUs 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 information used for the representation for both of the eyes and both of the eyebrows, while the six nodes in the output layer indicate the intensities of 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 mouth and nose information, while the 11 nodes in the output

layer represent the intensities of the recognised 11 lower facial AUs.

The training algorithms for both of the upper and lower AUs recognisers minimize the changes made to their corresponding network at each step in order to maintain both of the neural networks' generalization capabilities. 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. 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 posed test facial expression data inputs.

Moreover, psychological research has also laid foundations for the mapping between the AUs and emotions embedded in facial expressions (Ekman et al., 1971; Coan and Gottman, 2007). The Specific Affect Coding System (SPAFF) (Coan and Gottman, 2007) 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 1 based on the suggestion of the above employed Cohn–Kanade Database.

Table 1 shows some general guidance about physical manifest of emotional facial behaviours for the six basic emotions. However, affective facial expressions could be diverse and different from one person to another. Thus the training data set of the upper and lower AUs analysers has gathered diverse facial expressions from each subject and the FACS-coded images 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).

We also implement a supervised neural network based facial emotion recogniser to recognise neutral and the six basic emotions from facial expressions represented by the derived 17 AUs. This facial emotion recogniser also has a three-layer topology with one input, one hidden and one output layer. The emotion recogniser accepts the overall 17 upper and lower AUs as inputs and outputs the recognised neutral and the six basic emotions embedded in facial expressions. Therefore it has 17 nodes in the input layer and seven nodes respectively in the hidden and the output layers. The network topology of this facial emotion recogniser is provided in Fig. 6.

The FACS-coded peak frame emotional images from Cohn–Kanade DFAT-504 database and the FACS-coded emotional facial expressions posed by the real subjects are also employed for the training data construction for this neural network based facial emotion classifier. We have produced 23 training data for each emotional category. Overall the training set contains 161 emotional facial data for facial emotion recognition.

Table 1
The mapping between the physical cues represented by AUs and emotions provided by the Cohn–Kanade dataset.

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

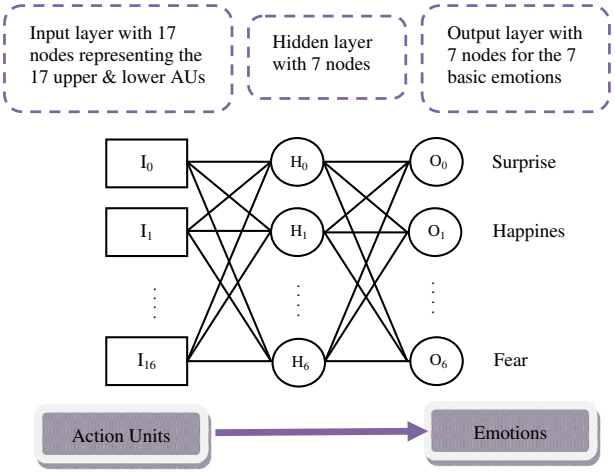

Fig. 6. The topology of the intelligent facial emotion recogniser.

Another five testing subjects are employed for the testing of the emotional facial behaviour recognition. 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. The real-time application 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 AUs analysers employ the corresponding derived facial data points as inputs and output the intensity scores for the selected 17 AUs associated with this facial expression. Secondly, these derived intensity values of the 17 AUs are subsequently used as inputs for the emotional facial behaviour classifier. This neural network facial emotion inference engine then outputs the detected emotion from this facial input.

NAO then conducts speech-based interaction with the testing subject to communicate back about the details of the facial emotion recognition results. Its speech synthesis engine is therefore activated to report the features of the upper and lower facial parts 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 if the recognised emotions from facial expressions 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 Facial Action Coding System are also used to remind the users about any particular emotional facial expression if help is needed. Otherwise, the users will freely demonstrate their emotional expressions based on their own interpretation during robot human interaction. Overall five testing subjects were involved in the testing and each testing subject showed five different types of facial behaviours for one specific emotion category. Detailed evaluation results for the 175 testing data are discussed in Section 6.

5. Topic detection using latent semantic analysis

After the detection of emotions from users' facial expressions, the robot will also start conversations with users by asking the reasons why they are experiencing a particular specific emotion. In order to deal with open-ended dialogue based interaction, we use a Latent Semantic Analysis based approach to topic detection. Such an approach is able to go beyond linguistic constraints and

provides a flexible and robust method for dialogue interpretation and management.

Latent Semantic Analysis (Landauer and Dumais, 2008) generally identifies relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. In order to compare the meanings or concepts behind the words, LSA maps both words and documents into a 'concept' space and performs comparison in this space. In detail, LSA assumed that there is some underlying latent semantic structure in the data which is partially obscured by the randomness of the word choice. This random choice of words also introduces noise into the word-concept relationship. LSA aims to find the smallest set of concepts that spans all the documents. It uses a statistical technique, called singular value decomposition, to estimate the hidden concept space and to remove the noise. This concept space associates syntactically different but semantically similar terms and documents. We use these transformed terms and documents in the concept space for retrieval rather than the original terms and documents.

In this research, we also employ the semantic vectors package (Widdows and Cohen, 2010) to perform LSA, analyze underlying relationships between documents and calculate their similarities. This package provides APIs for concept space creation. It applies concept mapping algorithms to term-document matrices using Apache Lucene, a high-performance, full-featured text search engine library implemented in Java (Widdows and Cohen, 2010). We integrate this package with the robot's platform to calculate the semantic similarities between users' inputs and training documents with clear discussion themes.

In order to calculate semantic similarities between test inputs with documents belonging to different topic categories, we have to collect some sample training documents with strong topic themes first. Personal articles from the Experience project (www.experienceproject.com) are borrowed to construct training documents. These articles belong to 12 discussion categories including Education, Family & Friends, Health & Wellness, Lifestyle & Style, Pets & Animals etc. Since we intend to perform discussion theme detection for those sensitive topics of young people that may influence their life dramatically, such as bullying, diseases and school life, we have extracted sample articles under the categories of Crohn's disease (five articles), school bullying (five articles), family care for children (five articles), food choice (three articles), school life including school uniform (10 short articles) and school lunch (10 short articles). Phrase and sentence level expressions implying 'disagreement' and 'suggestion' have also been gathered from several other articles published on the Experience website. Thus we have training documents with eight discussion themes including 'Crohn's disease (from which young people tend to suffer)', 'bullying', 'family care', 'food choice', 'school lunch', 'school uniform', 'suggestions' and 'disagreement'. The first six themes are sensitive and crucial discussion topics to teenagers, while the last two themes are intended to capture arguments expressed in multiple ways. Automatic recognition of metaphorical expressions often poses great challenges to linguistic processing systems. In order to detect a few frequently used affective metaphorical phenomena, we include four types of metaphorical examples published on the following website: http://knowgramming.com, in our training corpus. These include cooking, family, weather, and farm metaphors. We have also borrowed a group of 'Ideas as External Entities' metaphor examples from the ATT-Meta project databank (http://www.cs.bham.ac.uk/~jab/ATT-Meta/Databank/) to enrich the metaphor categories. Individual files are used to store each type of the metaphorical expressions, such as cooking_metaphor.txt, family_metaphor.txt and ideas_metaphor.txt etc. All the sample documents of the above 13 categories are regarded as training files and have been put under one directory for further analysis.

Moreover, NAO has embedded a speech recognition engine in its platform. When the human speaker provides a spoken utterance, the speech recognition engine performs the recognition and derives the recognised input utterance in a text form. This text-based input is then analysed by the semantic-based topic detection processing. The semantic-based processing performs recognition of discussion topics and metaphorical phenomena by calculating semantic similarity scores between this recovered user input and the above training corpus with 13 topic categories. For example, the semantic-based processing recognises the user input, "this disease is dragging me down", has the highest semantic similarities to the training documents under the categories of 'disease' (crohn_disease.txt: 0.81) and the mental metaphor corpus (ideas_metaphor.txt: 0.788). Moreover, the mental metaphor examples imply that the disease in this input has been regarded as being active outside of a living agent, which can carry out actions. The above semantic similarity results indicate that the input, "this disease is dragging me down", is more likely to contain topics of 'disease' and 'ideas metaphor". The semantic-based topic theme detection thus shows great potential in dealing with open-ended topic detection and metaphor recognition for those inputs extracted directly from published online articles.

Eighty pattern matching appraisal rules and knowledge base have been constructed to guide the response generation of NAO based on the detected emotions from facial expressions and the discussion topics of the users' inputs. An example rule is illustrated in the following.

(defrule example_rule (The identified topic theme is 'disease' or 'bullying' and the detected emotion from facial expression is 'anger') = > (obtain response from knowledge base)).

NAO normally responds to, on average, every *N*th speech by the human users, where *N* is a changeable parameter (currently usually set to 1 or 2). However, it also responds when the robot's name, NAO, is mentioned, and makes no response if it cannot detect anything useful both in the utterance and the subject's facial expressions it is responding to. As a future development we plan to have *N* dynamically adjustable according to how confident NAO is about what it has discerned in the speakers' facial expressions and the utterance at hand so that it is less likely to respond if it has less confidence. NAO makes a random response from several stored response candidates that are suitable for the affective quality it has discerned in the facial expressions.

In addition, NAO sometimes takes a portion of the user's input string as a sub-component of its response. For example:

<Human user> I'm your best mate.
<NAO> Why are you my best mate?

On the other hand, we also purposely put NAO in a daily life situation, conduct natural conversations with it and indicate spontaneous emotions in facial expressions to test its robustness. Some example interactions are provided in the following:

<Human user> You get lost! [Showing an angry facial expression; the detected topic is 'bullying']
<NAO> Hey, cool it folks.
. . .
<NAO> How are you doing?
<Human user> Not very good. [Showing a sad facial expression; the detected topic is 'school life']
<NAO> This does not look good. Let's play computer games together. Xbox game is my favourite. Come on. Cheer up.
. . .
<NAO> Have you heard that noise?

<Human user> Oh, someone breaks in the garden. [Showing a happy facial expression; the detected topic is 'family care']
<NAO> Oh, really. This doesn't sound good. Why are you laughing though?
<Human user> I'm just kidding. [Showing a happy facial expression; the detected topic is 'school life' and 'bullying']
<NAO> I knew it. I really hate you. I won't talk to you any more.

6. Evaluation

As mentioned in the training of the facial emotion recognition stage, NAO is able to recognise facial emotions from still database images although the current testing is mainly set up for the real-time affective facial expression recognition by real test subjects. This also shows potential to extend the testing of the system using images from the existing FACS-coded database. At the testing stage, as mentioned earlier, five testing subjects were involved in the testing and each testing subject showed five different types of facial behaviours for one specific emotion category. The 175 test data are employed for the evaluation of the overall intelligent facial emotion recognition system during real-time human robot interaction.

The experiments have been 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 mainly about. Then real-time facial expressions are posed by the testing subjects required by the robot. The recognised upper and lower facial actions and the embedded emotions in the facial expressions are eventually communicated back to the testing subjects by the robot's speech synthesis engine.

The neural network-based upper and lower facial AUs analysers are tested against the AU annotation provided by the certified human annotator using the 17 selected AUs and the AU-coded images extracted from the Cohn–Kanade database. The upper facial AU recognition achieved a 71.3% accuracy rate, while the lower facial feature detection achieved a 78.6% recognition rate. The Cohn–Kanade image database also discussed those Action Units shown during conversations tied to speech rhythm sometimes could be mis-regarded as indicators for emotional facial behaviours. For example, AU1, 2 and 4 tend to be used to express emphasis and requests for clarifications, and are often observed during conversations. They were regarded as the most frequent conversational signals by the Cohn–Kanade image database. In this research, we also noticed that upper AU 1, 2 and 4 are also frequently used to describe facial expressions cross several emotion categories for the training images extracted from the above image database. Such a training data set sometimes might confuse the system for the recognition of a few frequently used upper AUs cross several emotion categories. Also it seems that the employed lower AUs are more effective in describing emotional facial expressions and less overlapping with conversational signals compared to the upper AUs. This led to the fact that lower facial AUs were better recognised than the upper ones. Overall, the following AUs are reasonably recognised: 1, 2, 4, 6, 7, 9, 10, 12, 17, 23, 24, 25 and 26/27.

The neural network based facial emotion recogniser has also been tested against the affect labelling results provided by the testing subjects and the emotion annotation provided by the image database. The evaluation results indicate the neural network-based facial recogniser performed reasonably well with an accuracy rate of 71.3%. The detection results indicate that negative emotions such as anger (90%) and disgust (83%) have been well identified followed by the fear (65%) facial expression reasonably recognised, while sadness expressions pose the great challenge to the system.

The results also indicate that negative surprise and fearful facial behaviours have sometimes been misclassified as one another, while sad expressions are mostly miscategorised as angry behaviours. There is also strong resemblance between happy and positive surprise facial expressions demonstrated by the testing subjects during the testing. Although happy (80%) and positive surprise (77%) facial expressions are individually well recognised, these two positive emotional expressions are also sometimes misclassified as each other. Martinez and Du (2012) especially made some discussions about several different types of compound negative and positive surprise emotions. Thus their research may give some guidance on further finer classifications of these compound surprise facial emotions.

Moreover, the semantic-based topic detection processing is also evaluated at the testing stage. 200 sentences taken from both the Experience website (150 sentences) and the ATT-Meta databank (50 sentences) are used to evaluate the performance of LSA-based topic theme detection. The test sentences borrowed from the Experience website belong to the above mentioned topic categories: 'Crohn's disease', 'bullying', 'family care', 'food choice', 'school lunch', 'school uniform', 'suggestions' and 'disagreement'. Metaphorical examples are borrowed from the ATT-Meta databank targeting the following phenomena: cooking, family, weather, farm and ideas metaphors. Ten metaphorical examples are taken from each metaphorical category. Thus 200 sentences in total are used to evaluate the robustness and generalization abilities of the topic theme detection. The LSA-based topic detection achieves a 76% accuracy rate for the topic classification for the 150 test sentences borrowed from the Experience website and an accuracy rate of 83% for the recognition of the 50 metaphorical expressions. This LSA-based method proved to be effective in dealing with open-ended topic detection tasks.

7. Conclusions & future work

In this research, we have implemented *NN*-based upper and lower facial Action Units analysers and a *NN*-based facial emotion recogniser to detect emotions from real-time posed affective facial expressions. The system is also able to detect discussion topics from conversational inputs using a LSA-based topic theme detection processing. The *NN*-based upper and lower facial Action Units analysers are able to respectively derive six upper and 11 lower facial AUs with promising performances from motion-based facial representations. The upper AUs analyser achieves a 71.3% accuracy rate, while the lower facial analyser has a 78.6% accuracy rate. The NN-based facial emotion recogniser accepts the derived 17 AUs as inputs and shows promising performances for the decoding of seven basic emotions from facial expressions with a 71.3% accuracy rate. Most importantly, the intelligent facial emotion detection employs anatomical knowledge of facial Action Units as the bridge to link raw motion-based facial feature points automatically extracted by the robot with facial emotion recognition. The recognised 17 facial AUs also show great potential in interpreting a much wider category of human emotional facial behaviours. Moreover, the LSA-based topic detection is able to calculate semantic similarities between sentences to identify discussion topics beyond linguistic constraints. The LSA-based processing achieves an averaged accuracy rate of 79.5% for open-ended topic detection tasks. The overall development is dedicated to the NAO robot and integrated with its SDKs of Naoqi 1.12.5 C++ Linux 64 version. The work shows great potential in enabling the robot to deal with open-ended challenging robot human interaction.

In future work, compound cognitive emotion models will be explored in order to better recognise compound surprise facial emotions such as happy and angry surprise. The experiments show that NAO is also able to capture coloured facial image photos of the testing subjects in real-time. These image files will also be further

analysed using other techniques such as Gabor wavelets and Grey-level Co-occurrence Matrices to extract extra facial features in order to achieve finer classification of challenging emotion categories such as fearful and negative surprise facial expressions. Moreover, in future directions, we are also interested in using topic extraction to inform affect detection directly, e.g. the suggestion of a topic change indicating potential indifference to or un-interest in the current discussion theme. It will also ease the interaction if NAO is equipped with culturally related small talk behaviour. We also aim to incorporate each weak affect indicator embedded in semantic analysis and emotional facial expressions to draw more reliable affect interpretation in natural human robot interaction. We believe these are crucial aspects for the development of personalised intelligent agents/robots with social and emotion intelligence.

References

Afzal, S., Sezgin, T. M., Gao, G. & Robinson, P. (2009). Perception of emotional expressions in different representations using facial feature points. In *Proceedings of affective computing & intelligent interaction (ACII)*, Amsterdam.

Ammar, M. B., Neji, M., Alimi, A. M., & Gouarderes, G. (2010). The affective tutoring system. *Expert Systems with Applications, 37*(4), 3013–3023.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I. & Movellan, J. (2005). Recognizing facial expression: machine learning and application to spontaneous behaviour. In *Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR'05)* (pp. 568–573).

Coan, J. A., & Gottman, J. M. (2007). The specific affect coding system. In J. A. coan & J. J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment*. New York, NY: Oxford University Press [pp. 106–123].

Cohen, I., Looijeand, R. & Neerincx, M. A. (2011). 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).

Cohn, J. F., Reed, L. I., Ambadar, Z., Xiao, J. & Moriyama, T., (2004). Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *Proceedings of the IEEE international conferences systems, man, and cybernetics (SMC'04)* (Vol. 1, pp. 610–616).

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). Facial action coding system. *A Human Face*.

Ekman, P., Friesen, W. V., & Tomkins, S. S. (1971). Facial affect scoring technique: a field study. *Semiotica, 3*(1), 37–58.

Ekman, P., & Rosenberg, E. L. (2005). *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)* (second ed.). New York: Oxford University Press.

Ekman, P., Rosenberg, E. & Hager, J. (2013). Facial action coding system affect interpretation dictionary (FACSAID). <http://face-and-emotion.com/dataface/facsaid/description.jsp>. Accessed in Jan 2013.

Endrass, B., Rehm, M., & André, E. (2011). Planning small talk behavior with cultural influences for multiagent systems. *Computer Speech and Language, 25*(2), :158–174.

Ge, S. S., Samani, H. A., Ong, Y. H. J. & Hang, C. C. (2008). Active affective facial analysis for human–robot interaction. In *Proceedings of the 17th IEEE international symposium on robot and human interactive communication*, Germany, August 1–3.

Kaliouby, R., & Robinson, P. (2005). Real-time inference of complex mental states from facial expressions and head gestures. *Real-time vision for human–computer interaction*, 181–200.

Kanade, T., Cohn, J. F. & Tian, Y. (2000). Comprehensive database for facial expression analysis, In *Proceedings of the 4th IEEE international conference on automatic face and gesture recognition (FG'00)*, Grenoble, France (pp. 46–53).

Kappas, A. (2010). Smile when you read this, whether you like it or not: conceptual challenges to affect detection. *IEEE Transactions on Affective Computing, 1*(1), 38–41.

Kharat, G. U. & Dudul, S. V. (2008). Human emotion recognition system using optimally designed SVM with different facial feature extraction techniques, Ph.D. thesis, Anuradha Engineering College, Amravati University, India.

Landauer, T. K., & Dumais, S. (2008). Latent semantic analysis. *Scholarpedia, 3*(11), 4356.

Littlewort, G. C., Bartlett, M. S. & Lee, K. (2007). Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proceedings of the 9th ACM international conference on multimodal interfaces (ICMI'07)* (pp.15–21).

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn–Kanade dataset (CK+): a complete expression dataset for action unit and emotion-specified expression. In *Proceedings of the 3rd international workshop on CVPR for human communicative behavior analysis (CVPR4HB 2010)*, San Francisco, USA, pp. 94–101.

Martinez, A., & Du, S. (2012). A model of the perception of facial expressions of emotion by humans: research overview and perspectives. *Journal of Machine Learning Research*, 1589–1608.

McDuff, D., Kaliouby, R. E., Kassam, K. & Picard, R. W. (2011). Acume: a new visualization tool for understanding facial expression and gesture data. In *9th IEEE international conference on automatic face and gesture recognition (FG 2011)*, USA. pp 591–596.

Pantic, M. F., Valstar, M., Rademaker, R. & Maat, L. (2005). Webbased database for facial expression analysis. In *Proceedings of the IEEE international conference on multmedia and expo (ICME'05)*, Amsterdam, The Netherlands.

Ptaszynski, M., Dybala, P., Shi, W., Rzepka, R. & Araki, K. (2009). Towards context aware emotional intelligence in machines: computing contextual appropriateness of affective states. In *Proceeding of IJCAI*.

Rao, K. S., Saroj, V. K., Maity, S., & Koolagudi, S. G. (2011). Recognition of emotions from video using neural network models. *Expert Systems with Applications, 38*(10), 13181–13185.

Russell, J. A. (2003). Affect and the psychological construction of emotion. *Psychological Review, 110*, 145–172.

Schaaff, K. & Schultz, T. (2009). Towards an EEG-based emotion recognizer for humanoid robots. In *Proceedings of IEEE international symposium on robot and human interactive communication* (pp. 792–796).

Widdows, D. & Cohen, T. (2010). The semantic vectors package: new algorithms and public tools for distributional semantics. In *Proceedings of the IEEE international conference on semantic computing*.

Wong, J. J., & Cho, S. Y. (2009). A local experts organization model with application to face emotion recognition. *Expert Systems with Applications, 36*(1), 804–819.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1).

Zhang, L., & Barnden, J. (2012). Affect sensing using linguistic, semantic and cognitive cues in multi-threaded improvisational dialogue. *Cognitive Computation, 4*(4).

Zhang, L., Gillies, M. & Barnden, J. A. (2008). EMMA: an automated intelligent actor in E-drama. In *Proceedings of IUI, Spain* (pp. 409–412).

Zhang, L., Gillies, M., Dhaliwal, K., Gower, A., Robertson, D., & Crabtree, B. (2009). E-drama: facilitating online role-play using an ai actor and emotionally expressive characters. *International Journal of Artificial Intelligence in Education, 19*(1), 5–38.