Project Synopsis

Cover

+ Name of the project
+ Project Partners
  - Manas Kale :: 403037
  - Rohit Patankar :: 403053
  - Shubham Punekar :: 403061
  - Saurabh Shirodkar :: 403074
+ Company Name

  IBM
+ Name of the internal guide and co-guide
+ Name of the external guide

Certificate of completion (with MIT Logo)

Abstract

Abstract TODO Keywords : Human Computer Interaction, Machine Emotional Intelligence, Image Processing, Natural Language Processing, Speech and Audio Processing, Machine Learning, Affective Computing.

Acknowledgement

Contents

List of Figures

1 Problem Statement

Improve Human Computer Interaction with Machine Emotional Intelligence using Nao Robot, to recognize subjects based on their facial features and voice, analysis the facial features, speech text and the tone of speech to detect emotions with a weighted emotive score from former analyses and generate a context appropriate response on the robot.

2 Project literature survey

1. *Jeong-Sik Park, Gil-Jin Jang* "Implementation of Voice emotion Recognition for Interaction with Mobile Agent , ACM 2014"[1] The paper proposes a simple smartphone interface framework which consists of detection of human voice, extraction of emotional features and identification of an emotional state. Energy based approach for detection of human voice is selected in which if continues estimates of spectral energy of consecutive frames exceeds a pre-determined threshold, the region is regarded as starting point of voice signal. The pitch, log-energy, and "Mel-Frequency Cepstral Coefficient (MFCC)" are selected for extraction of emotional features which make a feature vector sequence. Acoustic features vectors extracted are analyzed and compared with patterns for each emotion type. Guassian Mixture Model is the classification algorithm used. This approach achieved 70.1% correctness within 1s response time. The future work suggested is applying proposed concepts for human-machine interaction in personal agent applications.

2. *Yu Gu, Eric Postma, Hai-Xing Lin* "Vocal Emotion Recognition using Log-Gabor Filters , ACM 2015"[2] The propsed work utilizes 2d Gabor filters in order to decompose the associated spectogram in order to perform a spectro-temporal analysis of affective vocalizations. Instead of including all potentially relevant features which leads of dimensionality problem and subsequent degradation of performance, the work uses "feature learning" in which relevant features are automatically obtained from raw speech signals. However this leads to considerable computational resouces. Hence, no. of features is kept to minimum. By performing analysis on local specto-temporal structure, the spectrogram is treated as an image and standard image processing is implemented. Comparative evaluation of MFCC and LPCC features, untuned and tuned Gabor filters and all above combinations is done. SVM is used as a classifier. The confusion matrix for performance using tuned Gabor Filters provide a maximum of 91.6% accuracy whereas a combination of acoustic features and Gabor filter provides 93.5% accuracy.

3. *Gloria Zen, Elisa Ricci, Nicu Sebe* "Unsupervised Domain Adaption for Personalized Facial Emotion Recognition, ACM 2014"[3] A personalization approach is proposed in which only unlabeled target-specific data are required. A new method to represent the source sample distribution based on only Support Vectors of source classifiers is proposed.

Regression framework is used to learn a mapping between a marginal distribution of the data points associated to a given person and the parameters of his/her personalized classifier which is represented by a set of Support Vectors of linear classifier in the source case and by all unlabeled data points in the target case.

4. *Ahmed Mustafa Mahmoud, Wan Haslina Hassan* "Determinism in Speech Pitch Relation to Emotions, ICIS 2009"[4] A deterministic rule-based text-to-speech emotional synthesis approach is proposed to generate emotional speech using semitonic interval-driven rules. Emotional speech samples are analyzed and intervals are extracted using praat tool. Objective evaluation compares sysnthesized voice to natural voice and calculates difference as an error function by considering mean square error as a measure of similarity. New emotional states may be defined using same proposed approach. Algorithms that integrate two or more emotional states may be combined to generate a variety of complex emotions.

5. /Nancy Semwal, Abhijeet Kumar, Sakthivel Narayanan/$_{\text{autom}}$ "Automatic Speech Emotion Detection using Multi-Domain Acoustic Feature Selection and Classification, IEEE 2015"[5] The proposed approach concentrated on determining emotions from speech signals. Various acoustic features such as energy, zero-crossing rate(ZCR), fundamental frequency, Mel Frequency Cepstral Coefficient are extractedfor short term, overlapping frames derived from the input signal. A feature vector for every utterance is then constructed by analyzing mean, median, etc. over all frames. Sequential Backward Selection is used with K-fold cross validation to select a subset of useful features. Detection of emotions is done by classifying respective features from the full candidate feature vectors into classes, using either a pre-trained SVM or a Linear Discriminant Analysis classifier. Accuracy of 80% was obtained when tested on EmoDB dataset.

6. *Lei Pang, Chong-Wah Ngo* "Multimodal Learning with Deep Boltzmann Machine for Emotion Prediction, ACM 2015"[6] In contrast to existing works which concentrate on either Audio, text or video, a joint density model is proposed over the space of multi-modal inputswith Deep Boltzmann Machine. The model is trained directly on user-generated Web videos without any labelling effort. Multiple layers of hidden units and multiple modalities make learning difficult, hence learning is split into 2 stages. First, each RBM component is pre

trained using greedy layerwise strategy. Then, learnt parameters are used to initialize the parameters of all layers in DBM and then the multimodal DBM is trained to finetune different modalities in a unified way. A major factor is that the deep architecture enlightens the possibility of discovering highly non-linear relationships between low-level features across different modalities. A performance improvement of 7.7% in classification accuracy is observed.

7. *Benjamin Guthier, Rajwa Alharthi, Rana Abaalkhail, Abdulmotaleb El Saddik* "Detection and Visualization of Emotions in an Affect-Aware City, ACM"lp In the proposed work, emotions are represented as four-dimensional vectors of pleasantness, arousal, dominance and unpredictability. In the training phase, emotion word hashtags in the messages are used as the ground-truth emotion contained in a message. A neural network is trained by using the presence of words, hashtags and emoticons in the message as features. During the live phase, these features are extracted from geo-tagged Twitter messages and given as input to neural-network. The detected emotions are aggregated over space and time and visualized on a map of the city.

8. *Jia-Ching Wang, Yu-Hao Chin, Bo-Wei Chen* "Speech Emotion Verification using Emotion Variance Modeling and Discriminant Scale-Frequency Maps"

9. *Lucile Bechade, Guillaume Dubuisson, Mohamed Sehili* "Behavioural and Emotional Spoken Cues Related to Mental States in Human-Robot Interaction"

10. *Fabien Ringeval, Shahin Amiriparian, Florian Eyben, Klaus Schrer* "Emotion Recognition in the Wild: Incorporating Voice and Lip Activity"

11. *Ali Yadollahi, Ameneh Gholipour Shahraki, Osmar R. Zaiane* "Current State of Text Sentiment Analysis from Opinion to Emotion Mining

12. *Jie Shen, Ognjen Rudovic, Shiyang Cheng, Maja Pantic* "Sentiment Apprehension in Human-Robot Interaction with NAO"

13. *Wei Jang, Wei Wang* "Face Detection and Recognition for Home Service Robots with End-To-End Deep Neural Networks, IEEE 2017"

14. *Rajesh K M, Naveenkumar M* "A Robust Method for face Recognition and Face Emotion Detection System using Support Vector Machines, IEEE 2016"

15. *Dario Bertero, Pascale Fung* "A First Look into Convolutional Neural Network for Speech Emotion Detection, IEEE 2017"

Literature Gap In general literature available today, numerous features have been developed, however the performance of classifiers is still limited, which is because of the fact that emotional states cannot be accurately distinguished by a well-defined set of discriminating features. Also, majority of work done towards emotion detection is focused on a single mode i.e. audio/ text/ video. There is limited practical work done with multimodal inputs.

3 Problem Definition

Knowing the emotional state of an individual can be crucial in determining what action is to be taken as a response. Recognizing the affective state of a human can be difficult for humans as well as computer systems. Many features can be considered such as voice samples, facial cues or even text written by the person to identify the emotional state of the individual.
The major focus of the project is improving human-machine interaction using the NAO robot. The robot will accept the input from the person periodically in the form of speech samples, comprising of voice and text as well as facial cues and will interpret the current emotional state of the person. Although our main focus is on humanizing the NAO robot and making it an ideal companion for old people, there are myriad of other uses that can be achieved; some of which are: development of an affect-aware city, Add security layer at public venues to detect malicious intent and deal with hostage situation effectively, measure response and ratings in focus groups (consumer response to commercials etc), wearables that help autistics discern emotion etc.

4 Scope of the problem

NAO robot will automatically and periodically analyze voice samples and facial cues in order to detect the emotional state of the person interacting with the robot. Depending on the emotions the person is feeling, the NAO robot will give an appropriate response. The response will be a combination of vocal response as well as physical gesture. This humane response will make the robot an ideal companion for old people. The robot will not be able to detect every single complex emotion, but will be limited to a subset of generalized emotions.

5 System architecture

5.1 High Level Design

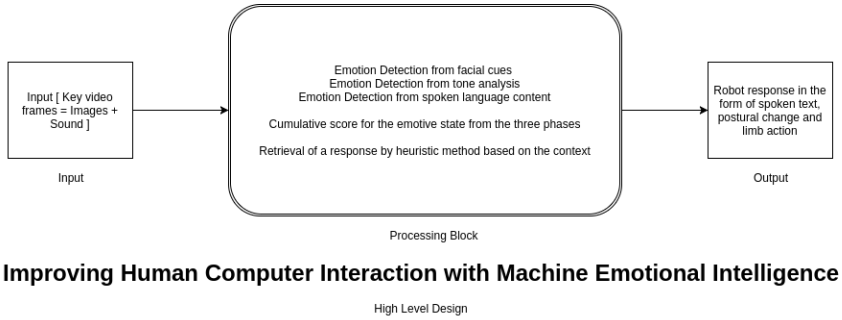

Figure 1: High Level Design

5.2 Low Level Design

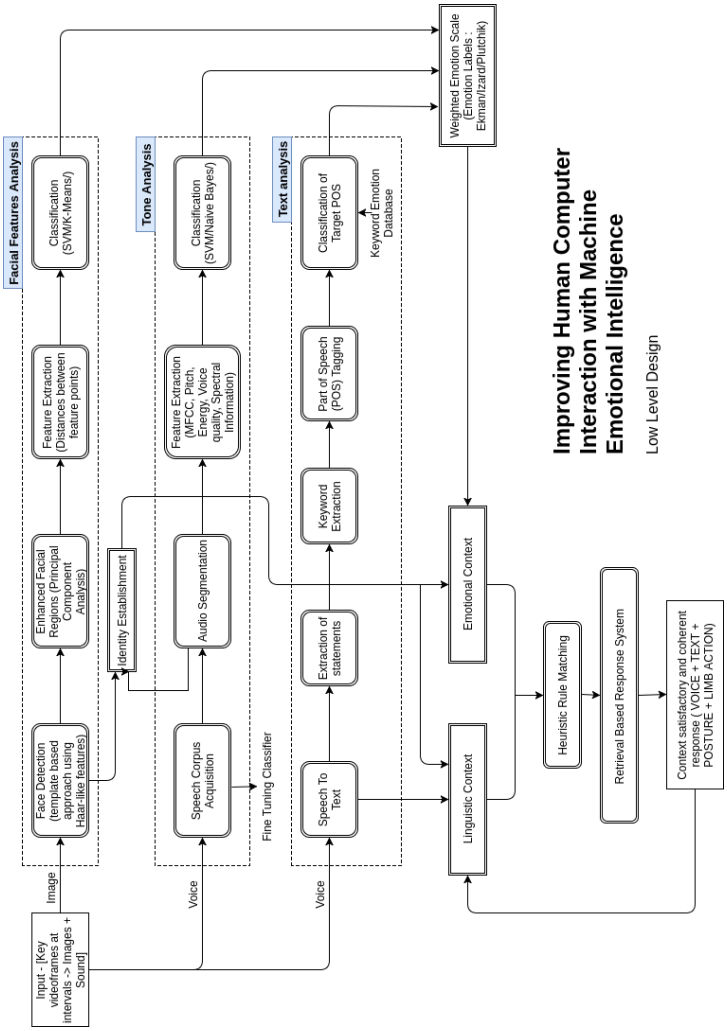

Figure 2: Low Level Design

6 Hardware and software requirements

6.1 Hardware Requirements

- NAO Robot : Softbank Robotics
  - Height : 58 centimeters
  - Wieght : 4.3 kg
  - Power Supply : Lithium battery providing 48.6 Wh
  - Degrees of freedom : 25
  - Autonomy : 90 minutes (active use)
  - CPU : Intel Atom @ 1.6 Ghz
  - Built-in OS : NAOqi 2.0 (linux-based)
  - Programming Languages : C++, Python, Java, MATLAB, Urbi, C, .NET
  - Sensors : Two HD Cameras, four microphones, sonar rangefinder, two infrared emitters and receivers, intertial board, nine tactile sensors, eight pressure sensors.
  - Connectivity : Ethernet, WiFi
- Server Requirements
  - RAM : 8 GB+ 1333/1600 Mhz
  - CPU : Intel Core (i5/i7 Family)
  - GPU : NVIDIA GPU Accelerator (GeForce Series 9/10 Family)
- Configuration for training classifiers :
  - RAM : 16 GB+ (DDR4 preferred)
  - NVIDIA Tesla GPU Accelerator (K40)
  - Intel Xeon Processor (E5/E7 Family)

6.2 Software Requirements

- Python packages (NPToolKit, python networking libs, etc)
- Continuum packages

7 TODO Feasibility study

7.1 Hardware feasibility

7.2 Software feasibility

Existing projects :

7.2.1 Emotion Recognition with facial detection

Emotive analytics : blend of psychology and technology. Although reductive, emotions clubbed into 7 main categories : Joy, Sadness, Anger, Fear, Surprise, Contempt, Disgust. For facial emotion detection, algorithms detect faces within photo or video, sense micro-expressions by analyzing the relationship between points on the face, based on curated databases compiled in aacdemic environments. **Sentiment analysis** processing software can analyze text to conclude if a given statement is generally positive or negative based on keywords and their valence index. **Sonic algorithms** analyze recorded speech for both tone and word content

Emotient For adverisement campaigns that want to track attention, engagement and sentiment from viewers. Provide RESTful Emotient Web API.

Affectiva Solution for massive scale engagement. SDKs and APIs offered for mobile developers.

EmoVu Facial detection product incorporates machine learning and micro-expression detection that allow accurate measurement of content's emotional engagement and effectiveness on their target audience. Desktop SDK, Mobile SDK and API for fine grained control provided by Eyeris. Other features offered by the platform are head position, tilt, eye tracking, eye open/close etc.

Nviso Specialise in emotion video analytics, using 3d facial imaging tech to monitor many different facial data points to produce likelihood for 7 main emotions. Aearder for smarter computing in 2013 by IBM.

Kairos Emotion Analysis API as Saas, coordinates detected in the input video that represent smiles, surprise, anger, dislike and drowsiness.

Project Oxford by Microsoft Catalogue of artificial intelligence APIs focussed on computer vision, speech and language analysis. Demo takes

a photo as an input and output is given in the form of JSON file, with detected faces and emotions of each, as a score between 0 to 1 for each of 8 emotions : anger, contempt, disgust, fear, happiness, neutral, sadness and surprise.

Face Reader by Noldus Used in academic sphere, Face Reader API is based on machine learning, dataset of 10,000 facial expression images. API uses 500 key facial points to analyze 6 basic facial expressions to analyse emotions, as well as gaze direction and head orientation.

SightCorp Facial Recognition Provider. Insight SDK tracks hundred of facial points, eye gaze tested in museum showcases and at TEDx Amsterdam.

SkyBiometry Cloud based face detection and recognition tool for detecting emotion in photos. Output is a percentage rate for moods : happy, sad, angry, surprised, disgusted, scared and neutral, in a given photo input.

Face++ Facial recognition tool that compares faces with stored faces, targeted for name tagging in photos in social networks. Determines if face is smiling or not. Provides a set of developer SDKs.

Imotions Biometric research platform providing software and hardware for monitoring many types of bodily cues. Imotion syncs with **Emotient's facial expression technology and adds extra layers to detect confusion and frustration.** Imotions API can monitor video live feeds to extrat valent, or can aggregate previously recorded videos to analyze for emotions. **Used by Harvard, Procter and Gamble, Yale, US Air Force**

CrowdEmotion API that uses facial recognition to detect the time series of the six universel emotions defined by Psychologist Paul Ekman (happniess, surprise, anger, disgust, fear and sadness). Analyses facial points in real-time video and respond with detailed visualizations.

FacioMetrics Founded at Carnegie Mellon University(CMU), provides SDKs for incorporating face tracking, pose and gaze tracking, and expression analysis. Can be tested using **Intraface iOS app**.

7.2.2 Text to Emotion

Sentiment analysis APIs that provide categorization or entity extraction. Following APIs specifically respond with an emotional summary given a

body of plain text.

Natural Language Processing use of machines to detect "natural" human interaction

Deep Linguistic Analysis examination of sentence structure, and relationship between keywords to derive sentiment

IBM Watson Powered by supercomputer IBM Watson, Tone Analyzer detects emotions tones, social propensities and writing styles from any length of plain text. **API can be forked on GitHub.** IBM also provides other cognitive computing tools.

Receptiviti Natural Language Personality Analytics API uses a process of target words and emotive categories to derive an emotion and personality from texts. Their Linguistic Inquiry and Word Count (LIWC) text analysis process used by IBM. Provides endpoints for REST API and SDKs in all major languages.

AlchemyAPI Determines relevance of keywords and their associated negative/positive connotations to get a sense of attitude or opinion. URL input can be given to recieve a grade of positive, mixed or negative overall sentiment. Overall sentiment evaluation for the document.

Bitext Text Analysis API is deep linguistic analysis tool. Can be used to analyse words, relations, sentences, structures and dependencies to extract bias with sentiment scoring functionality.

Mood Patrol Hosted on Mashape API marketplace, extracts emotions from text. It responds with fine grained adjectives that describe emotional tone based on Plutchik's 8 Basic Emotions.

Synesketch(opensource) Analyzes text for sentiment, converts emotional tone into visualizations. **Third-party apps constructed with Synesketch to recognize and visualize emotion from Tweets, speech, poetry and more.**

Tone API Quantifies emotional response for given content. Tool takes a body of text and analyzes for emotional breadth, intensity and comparison with other texts. **Possible application as a service for automating in-house research to optimize smart content publishing.**

Repustate API Repustate Sentiment Analysis process is based in linguistic theory, reviews cues from lemmatization, polarity, negations and parts of speech and more to reach informed sentiment from a text document.

7.2.3 Speech to Emotion

Speech recognition APIs are processed by other sentiment analysis APIs listed above, taking into consideration the **inflection of the speech**. Easy-to-consume web API that instantly recognize emotion from recorded voices are relatively rare. (Use cases : monitoring customer support centers, providing dispatch squads automated emotional intelligence)

Good Vibrations Good Vibrations API senses mood from recorded voice. API and SDK use universal biological signals to perform real time analysis of the user's emotion to sense stress, pleasure, or **disorder**. **EMOSpeech** is enterprise software to analyze emotion. "Audeering" software detects emotion, tone and gender in recorded voice.

Vokaturi Open Vokaturi SDK computes percent likelihoods for 5 emotive states : neutrality, happiness, sadness, anger and fear. (API has code samples for C and python)

8 TODO Design (UML Diagrams)

9 TODO Time-line analysis of the project

10 TODO Future scope

11 TODO Conclusion

References

[1] Jeong-Sik Park and Gil-Jin Jang. Implementation of voice emotion recognition for interaction with mobile agent. *HAI*, 2015.

[2] Yu Gu, Eric Postma, and Hai-Xiang Lin. Vocal emotion recognition with log-gabor filters. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15*, page nil, - 2015.

[3] Gloria Zen, Enver Sangineto, Elisa Ricci, and Nicu Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, page nil, - 2014.

[4] Ahmed Mustafa Mahmoud and Wan Haslina Hassan. Determinism in speech pitch relation to emotion. In *Proceedings of the 2nd International Conference on Interaction Sciences Information Technology, Culture and Human - ICIS '09*, page nil, - 2009.

[5] Nancy Semwal, Abhijeet Kumar, and Sakthivel Narayanan. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, page nil, 2 2017.

[6] Lei Pang and Chong-Wah Ngo. Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15*, page nil, - 2015.