Emotional Communications in Robots

White Paper by Matthew Walker Kings and Dave Catlin 23rd January 2006

Table of Contents

Introduction 1

Kismet 1

Robocasa 2

Sensory Input 2

Visual Input 2

Auditory Input 3

Touch Input 4

Environment Input 4

Input from the web 4

Direct input of information through keyboard etc. 4

Other Input Sensors 4

Sensory Output 4

Physical Output 4

Auditory Output 5

Touch Output 5

Output to the web 5

Direct output of information through a screen 5

Other Output 5

Responsiveness 5

Robosapien II 6

Roboraptor 6

Furby II 7

Cultural Considerations 7

Conclusion 7

References 8

Introduction

This white paper looks at the possibilities for emotional communication in the Serota project and relates these to existing research and knowledge in the field.

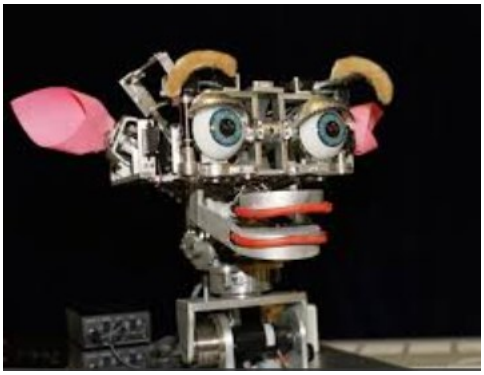
Key research work exists from a project called Kismet which was carried out by Dr Cynthia Breazeal At MIT in USA [1].

Other relevant work has been carried out as part of the Robocasa project at Waseda University in Tokyo [2].

Emotional communication relies on sensors in the robot to take input of the emotional signals sent by the human subject and actuators to output emotional signals from the robot to the human subject. The remainder of this paper discusses the input and output requirements of an emotionally responsive robot, the current state of the required technologies and cultural considerations.

Kismet

The Kistmet project ran from 1997 to 2003. The project is described as follows:

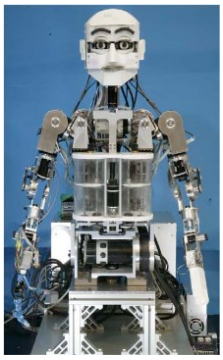


"The Sociable Machines Project develops an expressive anthropomorphic robot called Kismet that engages people in natural and expressive face-to-face interaction. Inspired by infant social development, psychology, ethology, and evolution, this work integrates theories and concepts from these diverse viewpoints to enable Kismet to enter into natural and intuitive social interaction with a human caregiver and to learn from them, reminiscent of parent-infant exchanges. To do this, Kismet perceives a variety of natural social cues from visual and auditory channels, and delivers social signals to the human caregiver through gaze direction, facial expression, body posture, and vocal babbles" [3].

Kistmet was very much a "lab" project and required a system comprising 15 computers and microprocessors running a mixture of operating systems. In answer to the question "Did your robot Kismet ever learn much from people?" Breazeal is quoted as saying:

"From an engineering standpoint, Kismet got more sophisticated. As we continued to add more abilities to the robot, it could interact with people in richer ways. And so, we learned a lot about how you could design a robot that communicated and responded to nonlinguistic cues; we learned how critical it got for more than language in an interaction - body language, gaze, physical responses, facial expressions. But I think we learned mostly about people from Kismet. Until it, and another robot built here at M.I.T., Cog, most robotics had little to do with people. Kismet's big triumph was that he was able to communicate a kind of emotion and sociability that humans did indeed respond to, in kind. The robot and the humans were in a kind of partnership for learning" [4].

Robocasa



Robocasa, a joint Japanese/Italian project at Waseda university in Japan, is a head and torso robot with 59 Degrees of Freedom aimed at expressing emotions. The project began in 1995 and is still in progress. The objective of the project is:

"We have been developing the Emotion Expression Humanoid Robots since 1995 in order to develop new mechanisms and functions for a humanoid robot having the ability to communicate naturally with a human by expressing human-like emotion" [2].

The Robocasa robot requires three standard PCs running Windows XP.

Sensory Input

In order to process a human's emotional cues a robot requires a variety of sensors:

Visual Input

Aims are as follows:

1. To isolate faces and read facial expression e.g. sadness, joy or anger.
2. To detect movement of human bodies and determine emotion through Body Language e.g. identify a human who is agitated, paying attention or asleep.
3. To infer emotional communication through non-specific movement in the visual environment e.g. a disrupted classroom environment.
4. To infer information by analysing colour.
5. To infer information by analysing light levels.

A camera can be used to provide visual input. A stereo pair of cameras can provide information about the distance of a subject from the robot. Four cameras, as used in the Kismet project, can provide wide angle and detailed stereo views.

Processing of images, particularly multiple images, is processor intensive and is only really feasible on a full-specification PC which means that camera images must be relayed from a mobile unit using a wireless connection.

There's some research going on body language / gesture recognition from camera captured images e.g. Intel Open Source Computer Vision Library but this appears to be a small research field at present [5]. It's far more common to use heart-rate sensors, skin resistance and pupil movement to assess this [6].

Similarly there is fairly reliable code available for identifying faces in an image but the ability to infer the emotional state of a person is a developing field [5]. Microsoft is looking into this in order to gauge the level of frustration of a user of their software. The most advanced technology in this area is in the CCTV/security area where known criminals can be picked out from images but this work is controlled by commercial companies and is not generally available [12].

The amount of movement in an image can be used to broadly guess at the amount of activity in the field of view. If the mobile robot houses the camera it must take into account that whenever it moves it will generate the appearance of movement in the camera image.

Colour and light levels are simple to analyse from a single camera image, for example the new RoboSapien II toy robot includes a camera and a small on-board chip which can detect red, green, blue or skin tone. The amount of emotional sensory input that this provides is quite limited.

Auditory Input

Aims are as follows:

1. Convert speech to text and determine emotion from content.
2. Determine emotion through tone of voice.

Conversion of text to speech is a mature field with market leaders Dragon Naturally Speaking [7] claiming up to 160 words per minute dictation speed with 99% accuracy after a few minutes training.

Dragon requires considerable computing power, memory and disk space and cannot be run at present on anything other than a full-specification PC. In addition it has to be trained to each user and will not understand words being spoken to it unless it "knows" the person speaking.

The voice recognition systems that are used in mobile phones do simple matching of audio files to what is being spoken and do not "understand" the words that are being spoken to them. However, mobile phone companies can already see the advantage of extending this to true speech to text so that text messages and emails can be written by voice alone and research is being carried out in this area [8].

Systems are available which can understand any person such as the Telephonetics[9] system used by the Odeon Cinema which claims 98.5% accuracy for any regional accent but this only has to be able to distinguish between the names of the 109 cinemas in the UK. Philips Vocon is another leader in this field and has products which they claim are speaker independent without training and which can run on embedded chips rather than desktop PCs[10].

In order to be able convert speech to text there is a need to remove background noise from the signal. The Kismet robot required a lapel microphone but this can also be done by using a pair of focussed stereo microphones so that background noise can be removed by subtracting phase difference. A notable user of this technique is the Philips DIMI home entertainment project [11].

Despite the high accuracy claims of the manufacturers speech recognition is very frustrating when it doesn't work 100%. Here's a comment from one disappointed buyer of Dragon Dictate from amazon.co.uk:

"I read the reviews and thought great I'll buy it … If you have a dialect (I am Welsh though I don't have a strong accent) then you will have to learn to adapt you speech and speak with gaps between the words but not slowing down the word itself. The program enters phrases which are quite random and it is very unpredictable. Just when you think you've got to grips with it, it all goes pear shaped!!! If you haven't got a lot of time to train the program then I'd practice increasing your typing speed instead of purchasing it."

In order to assess emotional state Kismet analyses the "spoken affective intent" of a voice using a system developed at MIT by the Spoken Language Systems Group. This takes into account the pitch and tone of a voice in addition to the word content.

If the multiple meanings of the phrase "watch it" are considered it's evident that words alone cannot convey emotional meaning and tone and intonation must also be considered.

Touch Input

A robot may be pushed, slapped, punched, scratched, kicked, stroked or cuddled. Distinguishing between these touches is important in assessing the emotional state of the person that the robot is interacting with.

Touch input can be read using sheets of multi-level force sensor material at various points on the robot's body. This subject is covered well in the Robocasa project documentation [2].

Environment Input

Various sensors are available for measuring environmental conditions such as temperature but few are indicators of emotional state. Tilt, speed of motion and proprioception (the position of the robot's body) may be useful in assessing the state of a person if they are interacting with the robot physically.

All of these types of sensor are easily interfaced and produce simple data.

Input from the web

The Internet gives the robot access to other computers and to all manner of information but again this is of limited use in assessing a person's emotional state. One area that may be of use is to look up someone's history once they have been identified.

Direct input of information through keyboard etc.

Although it's not an ideal way to communicate it's generally accepted, at present, that humans communicate with computers and the Internet by typing words on a keyboard and making selections with a mouse. Variations on this include making selections on a touch-screen and pressing buttons that are connected to the computer in other ways.

In order to give full interactive experience between a robot and a human in the early days of the project it may be necessary to resort to this method where other methods are not available.

Other Input Sensors

Body sensors such as heart-rate and skin moisture levels can provide valuable feedback about the emotional state of a subject but require sensors to be attached to the body.

Sensory Output

In order for the robot to communicate emotions back to a person various actuators must be used:

Physical Output

A human face is able to express six basic facial expressions: sadness, anger, joy, fear, disgust and surprise. The mechanics required to move a face on a robot in this way are quite complex.

In addition humans reinforce their facial expressions and verbal communication by moving their bodies and limbs -–body language.

There is no requirement for a robot to use the same facial expressions and body language as a human or an animal. If a robot were to spin three times clockwise it could indicate happiness. If it turned blue it could indicate dissatisfaction. These behaviours would need to be learned by those interacting with it.

Alternatively, simplified animal and human behaviours could be used as a starting point for interaction. For example a dog wags its tail when happy and humans have trembling limbs and raised eyelids when they are frightened.

The aim will be to give a range of emotional expressions whilst using simple actuators. For example the colours of parts of the robot's body could be changed through use of lights, video displays and smart materials.

Auditory Output

Again there is no need for a robot to use human speech. The popular robot toy Furby speaks in its own language and a dictionary is supplied so that this can be interpreted. Children who play with Furby quickly learn the "Furbish" language.

Pitch and tone can also be used to convey emotional state. The sounds that a baby makes do not form words but can be interpreted even when the baby is not visible.

Kismet does not attempt to produce words: "The robot's vocalization capabilities are generated through an articulatory synthesizer. The underlying software (DECtalk v4.5) is based on the Klatt synthesizer which models the physiological characteristics of the human articulatory tract. By adjusting the parameters of the synthesizer it is possible to convey speaker personality (Kismet sounds like a young child) as well as adding emotional qualities to synthesized speech (Cahn 1990) [4]."

In contrast speech is a very powerful method of communication and text to speech technology is mature, reliable and readily available for PC and embedded processor architectures.

Touch Output

Assuming the mechanics can be safe it may be possible for a robot to push cuddle or stroke a human in order to convey an emotional state.

Output to the web

A robot could communicate its emotional state to a PC or across the internet for monitoring purposes and to feed back into a person's history.

Direct output of information through a screen

Where no other method is available information could be displayed on a screen that's integrated into the robot. This would also be the only way for the robot to communicate by displaying pictures and video information.

Other Output

It would be possible for the robot to play music which is well researched as an emotional trigger and mood alteration device.

Responsiveness

Responsiveness is a key indicator of the "smartness" of the robot. A slow response gives the impression of a "dull witted" personality and people interacting with it will consequently have less regard for it.

The state of the art interactive robots in the field of toys are RoboSapien II and RoboRaptor from Wowwee but here are some responses from people who have bought it on amazon.co.uk:

Robosapien II

Reviewer: **A toy enthusiast** from Bradninch, Devon United Kingdom

"What a disappointment this is. From the hype and the sales descriptions I'd expected much - kind of like a little guy wandering around the home responding to controls and interacting with his environment.

In reality it cannot walk far or fast or even cope with a change in the thickness of the carpet, the batteries run down really quickly, so that it cannot walk at all after a matter of minutes and the hand movements are broken after only a couple of hours play time.

In a way it is interesting as an art object illustrating how poor robotic toys still are and what a way they have to go yet to fulfil the hopes we may have of them. I'm glad I bought it as a bit of fun for my middle aged self rather than as a present for one of the children as they could only have had a significant disappointment."

Roboraptor

"My children (aged 6 & 8) were unimpressed with this toy. It is undoubtedly a brilliant robot, but once it has walked about a bit, sniffed and growled a few times - what do you do with it? It doesn't encourage imaginative play and apart from the initial novelty is actually quite boring. The bite function is disappointing too - I thought it would actually pick things up in its mouth. Also it is supposed to detect obstacles in roam mode but it keeps getting stuck in corners! It can't walk on most of our floor surfaces either. All in all a waste of money."

Furby II

Conversely the Furby II to from Tiger Toys/Hasbro which, in my opinion, is the most successful interactive toy robot on the market. Again from amazon.co.uk here is some positive feedback about the emotional communication and projected personality:

"This is a great toy for adults and children alike. The educational value of it is it teaches children to be sociable and to speak. Also this toy is very fun."

Cultural Considerations

A video called 4 Sensations on the Robocasa web site shows the robot being slapped in the face and responding by looking surprised and moving its head away from the slap.

Smacking of children is illegal in several Scandinavian countries and is likely to become illegal soon in the UK. Even in a situation where the robot is likely to be subject to physical abuse such as a class of violent children the robot should not react to violence as this would be likely to encourage this behaviour.

Similarly, with reference to the Emotion Expression video on the Robocasa web page, responses for happiness, sadness etc are omni-cultural in facial expressions but are culture dependent in body language.

The emotional interaction of the SEROTA robot should aim to be suitable for all cultural situations but may need some level of customisation for each.

If the robot uses speech to convey emotion or interprets the emotional content of speech then customised language modules will be necessary. Even the differences between the same language spoken in different countries or regions of the same country will need to be accounted for.

Conclusion

Simple communication from a person to a robot is straightforward but assessment of the emotional state of a person through sound and image is complex and will require considerable research work.

Communication of emotional state from a robot to a person is an established field. For the Serota project it will be necessary to define which methods of communication will be used which will be dependent on parameters such as cost.

Whatever method of communication is selected it must be implemented successfully and in a responsive way otherwise the robot will not be judged successful. For example, if a camera is fitted it is essential that the robot responds quickly and accurately to visual stimulus.

The degree of cultural customisation required depends upon the methods of communication that are selected.

References

[1] Kismet Overview

http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html

[2] Robocasa Overview

http://www.takanishi.mech.waseda.ac.jp/research/eyes/index.htm

[3] Kismet Interview with Breazeal

http://www.temple.edu/ispr/examples/ex03_06_10.html

[4] Kismet Hardware Design

http://www.ai.mit.edu/projects/sociable/baby-bits.html

[5] Intel OpenCV library

http://www.intel.com/technology/computing/opencv/overview.htm

[6] Assessing emotional state
http://www-personal.engin.umd.umich.edu/~xylum/ResearchIntroduction/Review-Affective%20State%20Assessment.htm

[7] Dragon Naturally Speaking

http://www.scansoft.co.uk/naturallyspeaking/

[8] Mobile Phone True Speech Recognition

http://www.wirelessnewsfactor.com/story.xhtml?story_title=Voice_Recognition_Matures_for_Mobile_Phones&story_id=40491

[9] Telephonics Cinema Speech Recognition

http://www.telephonetics.co.uk/telephonetics/uploads/pressreleases/200112HarryPotter-FINALAPPROVED.pdf

[10] Philips Vocon Speech Recognition

http://www.semiconductors.philips.com/acrobat_download/other/news/infocus/esc_conference/VoCon.pdf

[11] Philips Dimi Overview

http://www.research.philips.com/password/archive/23/downloads/pw23_personaltouch_14.pdf

[12] Image Processing to Detect Criminals

http://www.futurepundit.com/archives/000609.html