MIRA: A Learning Multimodal Interactive Robot Agent

John C. Murray[1], Stefan Wermter[2], Michael Knowles[2]
Hybrid Intelligent Systems,
School of Computing and Technology,
University of Sunderland, SR6 0DD, UK.
john@jcmurray.com[1], stefan.wermter, michael.knowles{@sunderland.ac.uk}[2]

Abstract

In this paper we present a robotic head MIRA (Multimodal Interactive Robot Agent) which has been developed for studying the learning of human robot interaction and improving our understanding of human robot interaction techniques. In this paper we focus on two main aspects of the system; first, we describe how the robot head learns to recognise faces for supporting the interaction process between a human and MIRA. Second, we show how MIRA can learn to identify sound sources of interest and attend to the source location improving the social interaction effect. We propose that there is substantial potential for learning visual and auditory features in order to increase adaptability and robustness of robotic heads.

1. Introduction

The development of robotic heads has recently been driven by the desire to create more socially interactive robots [4]. Since communication between humans relies heavily on both visual and verbal information [11, 2, 10] some approaches have addressed the combination of these modalities. However, in order to allow better robot-human interaction, it is not only necessary for robots to communicate but also to automatically adapt based on visual and auditory input. This automatic adaptation and learning is particularly important as it allows the human to feel that the interaction with the robot is being conducted in a natural manner [4].

There are many socially interactive robots in use, for example, Minerva [9] developed at Carnegie Mellon University is a tour guide robot that takes visitors around a museum. However, Minerva's learning, language and sound localisation capabilities are restricted in terms of natural social interaction. The University of Freiburg have developed an interactive robot called Fritz [1] but again learning and sound localisation is not being used to support visual recognition.

In this paper we describe an approach to learning facial recognition and sound-source localisation in order to support human-robot interaction in a more natural human-like manner, drawing on inspiration from biological systems. Sound-source localisation allows MIRA to detect and then orientate towards a speaker wishing to gain MIRA's attention. The facial recognition gives MIRA the ability to either learn a new face or to recognise a previously learnt face. We argue that both these learned auditory and visual capabilities are essential for developing a socially interactive robot. While there has been substantial research in facial detection [1] and some research in sound localisation [9] our focus in this paper is not on replicating this research performance but on focusing on integrating both tasks in one robot head as well as allowing for automatic adaptation and learning of these tasks as much as possible.

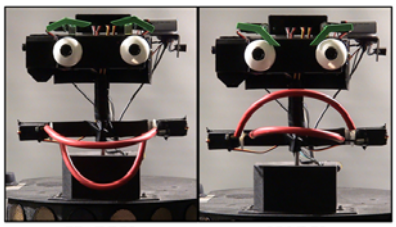

Figure 1. MIRA's emotional expressions.

2. Facial Recognition

MIRA (see Fig. 1) is designed to be capable of tracking and recognising human faces. The method used for initial detection of a face is based on Viola's [10] rapid object detection with improvements made by Lienhart [6]. In order to be able to detect a face, or any desired object of interest, specially trained classifiers are used. These classifiers are trained as described in [10] using a sample set of images that contain the particular feature we wish to detect; in our case this was a collection of face images.

We are aware that there are other more complex and accurate methods of face detection using such methods

as principle component analysis [7]. However, within this paper, our focus is on the multimodal interaction between auditory and visual information for human-robot interaction. The particular classifier used in our system is a cascade of boosted classifiers working with Haar-like features [6]. Haar-like features, instead of using the actual value of individual pixels, use the change in contrast values between adjacent rectangular groups of pixels. Contrast differences of selected groups of pixels are used to determine relative light and dark areas of a sample region within the image. Two or three adjacent groups of pixels with a relative contrast variance then form a Haar-like feature [6].

2.1 Facial Feature Extraction

In order to be able to learn and therefore, recognise a face, MIRA needs to have some mechanism of collecting facial information. For this we use the relative positions of the significant elements present on all faces: eyes, nose and mouth. It is therefore necessary to train several independent classifiers each responsible for the detection of a feature we wish to detect within the face. For our system, four classifiers were trained to resemble all the required features: a human face, eye, nose and mouth respectively Table 1 shows the number of images per training set.

Our aim was to study images taken under various conditions and with various cameras in order to build robustness into our learning systems. Initially we used the NIST FERET face recognition database [8]. However, the acquisition of these images is very specific, taken with the same lighting conditions, at the same distance from the camera and the same orientation. Therefore, we also collected training images using Google's image search tool providing the search criteria 'eye', 'eyes', 'nose' and 'mouth'.

Once the classifiers are trained they are applied to the images captured by MIRA's camera and processed accordingly. MIRA waits for an instruction to start interacting which can come in the form of a user saying "Hello", "MIRA" or a combination of the two. MIRA then orientates to the direction of the sound and begins to capture images. Each of the captured images is sent to the face classifier to see if the image contains an actual face. This process is fairly efficient, with MIRA being able to analyse an image for a face in less than 80ms allowing a frame rate of 12fps to be achieved.

If a face is detected it is extracted from the rest of the image and normalised by resizing it to 250px in width, maintaining a ratio of 1:1. It is then passed to the second stage of processing. This normalising ensures the faces are always the same size (in pixels) when presented to the feature extraction stage and this helps to prevent the system from classifying the same face as different people when viewed at different distances. However, if a face is more than three meters away the resolution of the image decrease too much to recognise. Fig. 2 shows the scene capture with the detected face (inserted). The bounding box in this image is displayed to show which section of the image the classifier extracts as the detected face, and is performed automatically by the face detection software.

Figure 2. Captured scene and face detection.

Table 1. Number of Training Images

Feature	Positive Samples	Negative Samples
Face	1,000	1,000
Eyes	500	100
Nose	500	100
Mouth	500	100

The face is split into three parts, as shown in Fig. 3, and each segment applied to the appropriate classifier Q1 & Q2 to the eye classifier and Q3 to the nose and mouth classifiers. Fig. 3 shows how the face is split into the separate parts for feature and data extraction. Once the features have been extracted from the image segments, unique feature data is created that is used to identify the face on future views. For this purpose we use the Euclidean distance of the features within the original face image. We find this is a useful measure of identifying different people due to the variance in the structure of people's faces. The first feature data we compute is the distance between the two eyes. In order to acquire this we need to have a reference point for the eye. We use the centre of gravity (COG) of the eye image extracted by the classifier for this. Fig. 4 shows the algorithm used for determining the COG.

Once the feature data has been extracted from the image, it is used to train a neural network for later facial recognition. A neural network provides some generalisation advantages over the use of a standard lookup table. The most important advantage is an artificial neural network's (ANN) ability to adapt to noise within the input data, which is very important in our deployment of MIRA in real world situations when people will be exposed under difference lighting conditions, rotations and distances.

2.2 Training the Neural Network

We use a feed-forward multi-layer perceptron (MLP) network with the following empirically determined architecture: four input units, one hidden layer containing 20 hidden units and ten output units. Each of the input units receives their activation from one of the extracted facial features. The feature data extracted from the facial image is normalised to a value between 0, 1 as determined by,

$$I_n = \left(\frac{1 + (x - A)}{B - A} \right) - 1 \qquad (1)$$

where I_n is the normalised value of the particular feature data that is input into the ANN, x is the value of the feature data in question, A is the minimum possible value of the feature data, and B is the maximum possible value. The ten output units represent separate faces that have been learnt by the network. With this particular architecture and scenario, the system is capable of learning a restricted number of faces; however, we feel a small number of faces (e.g. 10s rather than 1000s) are sufficient within a domestic scenario where there would be few people to be recognised, for instance in a single home environment.

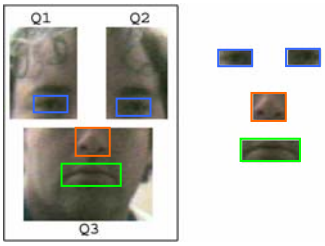

Figure 3. Individual face segments.

```
COG_X = COG_X + (I*x)
COG_Y = COG_Y + (I*y)
Total = Total + I
    where I = (R+G+B)/3, x, y is the current pixel location.
COG_X = COG_X/Total
COG_Y = COG_Y/Total
    results in the final x, y location of the COG.
```

Figure 4. The algorithm for calculating COG

To determine the capability of the system in face recognition the network was trained on ten different faces presented to MIRA. To gauge how the system would later respond to those same faces when they are presented at varying rotations. Each face was presented 20 times at each angle of rotation and the average recognition result taken. Table 2 gives the results of the system on recognising ten separate faces that the system has previously learnt. The faces were initially presented to the system with a rotation of 0° for the purpose of learning. Within the table a 'Y' indicates a successful recognition of ≥ 75% over the total number of trials, an 'N' represents < 75% whilst a '!' represents a misclassification of ≥ 50%.

These results show that the system provides an accuracy of 100% when the face is shown to the system with a rotation of 0°. In line with human performance, as the rotation increases, the accuracy of the system decreases to a point where at a rotation of ±20° the system only gives a 15% accuracy rate. However, the system is fairly reliable with the best results in the range of ±10°. Currently our system does not apply an in-plane or out-of-plane rotation on the face image.

Table 2. Recognition Rates of Faces

Face	0°	10°	20°	30°	-10°	-20°	-30°
1	Y	Y	N	N	Y	N	N
2	Y	Y	N	N	!	Y	N
3	Y	!	N	N	!	!	N
4	Y	Y	!	N	!	N	N
5	Y	Y	N	N	Y	N	N
6	Y	Y	N	N	Y	Y	N
7	Y	N	N	N	N	Y	N
8	Y	!	!	N	Y	N	N
9	Y	Y	N	N	N	!	N
10	Y	Y	N	N	N	!	N

3. Sound Source Localisation

MIRA is also capable of sound-source localisation, allowing the head to orientate itself to face the direction of a sound-source of interest, i.e. a person trying to attract MIRA's attention. This capability plays an important role for robot-human interaction: if a user wishes to attract the attention of MIRA when the attention/gaze of the robot is elsewhere, then the natural reaction would be to call the robots name "MIRA". The robot would then respond to this by determining the direction of the sound-source and turning to face the source.

The mammalian auditory system is very adept at localising sound-sources within an acoustically cluttered environment. Therefore, inspiration for the acoustic model developed is taken from that of the mammalian central auditory system (CAS). Within the CAS several cues are used to determine the direction of a sound-source. These include the Interaural Time Difference (ITD) and the Interaural Phase Difference (IPD) as two cues for azimuth estimation. The ITD is the delay between the sound signal arriving at the left and right ears, whereas the IPD is the phase difference of the two signals at the ears. There is evidence as shown by Licklider's triplex model [5] suggesting that cross-correlation exists in biology.

The localisation capabilities provided on MIRA are implemented using the IPD cue, and taken as the phase difference in the signals detected by the microphones. The two signals ($g(t)$ and $h(t)$) phase difference are calculated using a method known as cross-correlation [11]. This involves taking the signals $g(t)$ and $h(t)$ and, using a sliding window approach, calculating the phase difference by firstly offsetting the two signals and then

sliding them across each other, computing the product of the values at each point. This produces a correlation vector ***C*** (sizeof (g(t))*2)-1. The maximum value within this vector corresponds to the maximum point of similarity between the two signals. Using this value and its corresponding location within the correlation vector we can use the ITD cue to calculate the time difference between the signals arriving at the left then right microphone and thus determine the azimuth angle of the sound-source. We calculate this angle using the following equation.

$$\vartheta = Sin^{-1}\frac{a}{c} = Sin^{-1}\frac{(\Delta \times \sigma) \times c_{air}}{c} \qquad (2)$$

where, σ is the lag between *g(t)* and *h(t)* determined by correlation vector ***C***, Δ is the sample time increment determined by the sound card sample rate, i.e. 1/44100 = 22.7μs and ***θ*** is the angle of incidence, with the speed of sound being taken as 348 m/s at 24°C.

$$\vartheta = Sin^{-1}\frac{(22.7\mu s \times -17) \times 348m/s}{0.15m} \qquad (3)$$

$$\vartheta = -63.4^{o}$$

Therefore, using Eq. 3, when the system detects a sound of interest the azimuth angle of the source is calculated and MIRA orientates towards the direction. This orienting allows the speaker's face to come into view and for the MIRA to see if this is a previously recognised person or someone new using the visual recognition process. The performance of the system gives an accuracy of ±2° with the response more accurate in the range of 0° to ±45°. See Fig. 5.

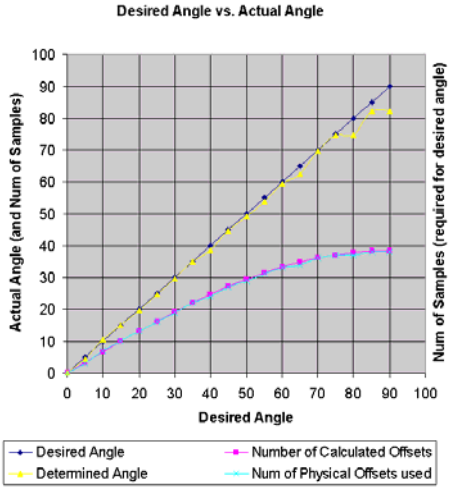

Figure 5. Accuracy of the sound source localisation

4. Discussion and Conclusion

We have shown the development of a multimodal learning robot agent MIRA. The results presented for the facial recognition show that the system is capable of learning a person's face and using this knowledge to recognise that person on successive presentations. The system we have presented also demonstrates the use of simple interactive behaviour using basic verbal expressions. Emotions allow the robot to convey more information in a human-like manner than is possible with just the verbal interaction aspects of the system.

In this paper we have described, for this first time, the overall architecture and first set of experiments with MIRA. Future work will focus on improving the recognition rate of the system by integrating transformations on the detected face images. Furthermore we consider scaling up the system to create a dynamically expandable learning system which can grow and develop with experience [3]. http://www.his.sunderland.ac.uk/projects/robot_head.htm shows MIRA in action.

5. Acknowledgements

This research has been partially supported by the MICRAM EPSRC grant EP/D055466/1 and partially by EU project NestCom under NEST-043374.

References

[1] Bennewitz M., Faber F., Joho D., and Behnke S., "Fritz – A Humanoid Communication Robot", Accepted for 16thd IEEE International Symposium on Robot & Human Interactive Communication (RO-MAN), Jeju Island, Korea, August 2007.

[2] Chen H.Y., Huang C.L., and Fu C.M., "Hybrid-boost learning for multi-pose face detection and facial expression recognition", Pattern Recognition, In Press, Accepted Manuscript, 2007.

[3] Hung C., and Wermter S., "A Dynamic Adaptive Self-Organising Hybrid Model for Text Clustering", The 3rd IEEE International Conference on Data Mining, pp. 75-82, Melbourne, USA.

[4] Kerepesia A., Kubinyib E., Jonssonc G.K., Magnussonc M.S., and Miklósib A., "Behavioural comparison of human–animal (dog) and human–robot (AIBO) interactions", Behavioural Processes, Vol. 73, Issue 1, Pages 92-99, July 2006.

[5] Licklider J.C.R., 'Three auditory theories', In: Psychology: A Study of a Science, ed Koch, E. S., Study 1, Vol. 1, New York: McGraw-Hill, pp. 41 – 144, 1956.

[6] Lienhart R., and Maydt J., "An Extended Set of Haar-like Features for Rapid Object Detection" IEEE ICIP, Vol. 1, pp. 900-903. 2002

[7] Menser B., and Muller F., "Face detection in color images using principal components analysis", Image Processing and Its Applications, Seventh International Conference on (Conf. Publ. No. 465), Vol. 2, Issue , pp. 620-624, 1999.

[8] Phillips P.J., Moon H., Rauss P.J., and Rizvi S., "The FERET evaluation methodology for face recognition algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, October 2000.

[9] Thrun S., Bennewitz M., Burgard W., Cremers A.B., Dellaert F., Fox D., Hähnel D., Rosenberg C., Roy N., Schulte J., and Schulz D., "MINERVA: A Tour-Guide Robot That Learns", Lecture Notes in Computer Science, Volume 1701/1999, Springer / Berlin / Heidelberg, page 696, 1999.

[10] Viola P., and Jones M.J., "Rapid Object Detection using a Boosted Cascade of Simple Features" IEEE CVPR, 2001.

[11] Weisstein E.W., "Cross-Correlation Theorem." From MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/Cross-CorrelationTheorem.html