Eye Contact as a Key Component in Human Robot Interaction

Shaun Press
shaun.press@anu.edu.au

Jason Saragih
jason.saragih@anu.edu.au

Jason Chen
jason.chen@anu.edu.au

Robotic Systems Lab, RSISE
College of Engineering and Computer Science
Australian National University
Canberra, Australia

Abstract

Eye contact is a key communication prompt in human-to-human interaction involving conversation, especially in large groups. In this paper we report on a project to construct a robotic exhibit at Questacon, the Australian National Science and Technology Museum. This robot will seek to interact with visitors to Questacon, including a conversational interface. This paper reports specifically on an eye-contact behaviour that we have implemented on the robot to couple with this conversational interface.

1   Introduction

When engaged in conversation in large groups, humans use eye-contact as a key prompt as to who conversation is being directed at. Subsequently, a good strategy for a robot to utilise when directing speech to a particular human in a large group is also to make "eye contact".

In this paper we report on a robot being constructed jointly by the Australian National University and Questacon. Questacon is the Science and Technology Museum located in Canberra, Australia, and its role is "To achieve its aim of promoting greater understanding and awareness of science and technology within the community, Questacon is committed to making an experience with science and technology fun, educational and interactive. Questacon has over 200 exhibits and approximately 300 000 people visit the Centre in Canberra each year." [1].

The robot being designed will be situated in the entry foyer of Questacon's exhibition centre in Canberra and will act as an exhibit to the public designed to

Figure 1: Questacon Robot - Artists Impression

---

[1] From "What's a Questacon?" http://www.questacon.edu.au

both assist, and to demonstrate practical robotic technology. One of Questacon's requirements for the exhibit is that it communicate verbally with humans. We report on the details of this verbal interface in [Press *et al.*, 2006]. Clearly there will be groups of visitors congregating around the robot in normal operation. In this paper we we report on a system that we have designed that selects and tracks a face in the group so that verbal output can be directed as appropriate. We intend to combine this system with a microphone array system that can detect the direction of audio. Although it is not implemented yet, the final aim will be to have a microphone array detect the general location of the speaker and steer the cameras to look for a face in this direction.

2 Questacon Robotic Project

The hardware platform for this paper is the Questacon Robot (see Figure 1. The robots internal structure was originally designed by past lab member, David Austin, and was constructed by Questacon. The exterior was designed and is currently being constructed by Questacon. The robot has a large number of sensors, including Laser Scanner, Infra-red sensors, bump sensors, active stereo vision head, microphones. For the first stage of the robots implementation, the robot will serve as a demonstration of various robot technologies. These technologies include speech recognition and response, face recognition and tracking, object detection, and localisation and mapping. For the robot to successfully demonstrate these technologies, it must engage the user in an interactive manner.

The robot will interact with users using 3 different systems. The first method is via the active head face tracking system. This system will select a face from a set of people standing in front of the cameras, and will follow the face using a combination of face recognition and skin colour tracking techniques. This interaction will last as long as the user is looking at the robot.

The second system is via a speech recognition and response system. Users will be able to speak to the robot and receive information in spoken form. This system uses a combination of grammar based recognition systems, as well as a general dictation recognition mode. Using both these system allows the robot to provide specific information upon request, as well as engaging the user via more general conversation.

The third system is a touch screen system mounted on the front of the robot. The touch screen will provide the user with various pieces of information concerning the operating state of the robot, as well as allowing the user to control which information is being presented.

In this paper we report on the face detection and tracking system.

3 Active Vision Face Tracking

The Active Vision Face Tracking system combines both facial detection (using the Viola-Jones face detection algorithm [Viola and Jones, 2001]) with the Camshift skin colour tracking algorithm [Bradski, 1998]. The tracking system detects and tracks faces using the cameras mounted on a CEDAR Head [Truong *et al.*, 2000] and moves the cameras to keep them focussed on the subject. The system continues to follow the track the subject for the period of time the subject remains in view of both cameras, and the subject demonstrates an interest in the robot by looking at the cameras. This process is shown in Figure 3

3.1 Vision Hardware

The vision system is a stereo camera system mounted on a CEDAR Head . The CEDAR Head is a cable drive
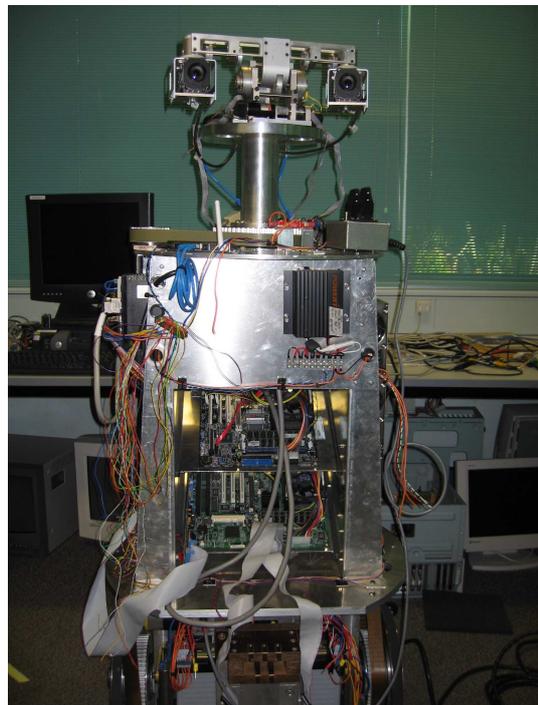

Figure 2: Questacon Robot with Active Head

system consisting of 3 independent motors. Each camera has a motor which allows them to pan left and right, while the third motor provides tilt control. The cameras rotate independently on there own vertical axis, with a tilt control along the horizontal axis. The entire CeDAR Head is then mounted on a motorised shaft, which simulates a human neck.

The motors receive their control signals from a Servo-to-go card attached to the robots motor control computer. The servo-to-go card outputs voltages in the range of +- 5 volts to each of the motor controllers. These voltages are translated into movement in the positive and negative directions. Feedback from the motors is provided by an optical encoder system which produces digital input back to the Servo-to-go card.

3.2 Vision Processing Software

The modules to handle face recognition is part of Daves Robotic Operating System (DROS http://www.dros.org), an implementation of Dca (Distributed Control Architecture for Robots [Petersson *et al.*, 2001]). The code itself is written in C++ and utilises both code written specifically for DROS as well as utilising the OpenCV vision processing and graphics library (http://www.intel.com/technology/computing/opencv/index.htm) The system was designed and implemented with the use of open source software. This allows the system to be developed further by other researchers and allows the current system developers to take advantage future developments in the vision processing software.

3.3 Detect Faces

For each frame returned from both cameras, the Viola-Jones Face detection algorithm is run across the image and a set of likely faces are returned.

The Viola-Jones detector is perhaps one of the most efficient general-purpose object detector to date. The method uses a cascade of boosted tree classifiers as a statistical model. By virtue of the ada-boost method, each level of the cascade is trained to give a very low false negative rate using a large sample set of face and non-face examples. All false-positives in levels lower in the cascade hierarchy are used to train the later ones such that the complexity of the discriminative model increases as we ascend the cascade. Nonetheless, most image regions which do not contain a face will be rejected at lower levels of the cascade, resulting in an efficient on-line evaluation. Furthermore, as the models at each cascade level

are built from simple haar-like features, they are very efficient to evaluate using the integral image, and hence, an exhaustive search at all locations and scales in the image can generally be performed at pseudo-real time.

One of the major drawbacks of the Viola-Jones method is that it requires a very large database of the object class and exhibits extensive training times (the original Viola-Jones formulation takes 2 weeks to train!). We use the pre-trained statistical model freely available with OpenCV as it works well enough.
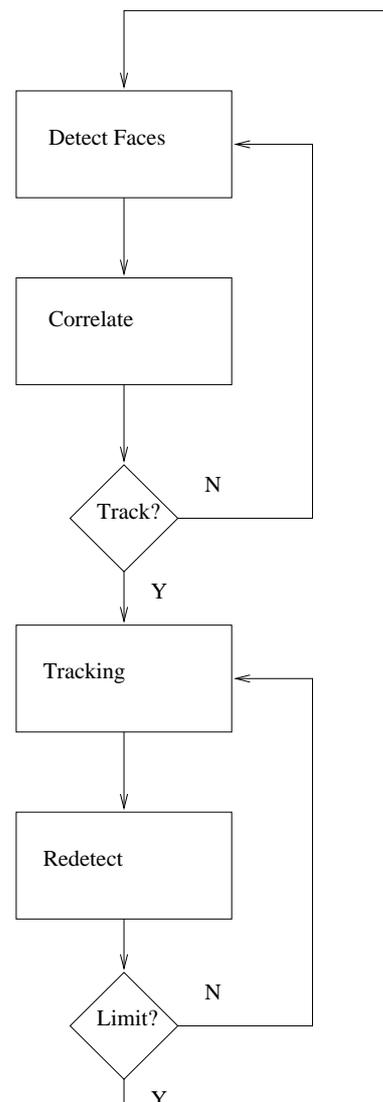

Figure 3: Flowchart of Detecting and Tracking States

3.4 Correlation

The faces detected in each frame are independent of each other and therefore a correlation process is required. For each face in the left camera, and matching process is carried out for all faces in the right camera. The first part of this process is the elimination of faces that could not match the face in the left scene. As the cameras are mounted parallel to the surface that the robot stands on, identical faces in each camera will be of equal height. Therefore any face that falls outside a predetermined height range of the left face will be immediately discarded. The remaining faces are then matched using similarity in colour and geometry. As the face detection algorithm returns a bounding box for each face, identical faces should return bounding boxes of a similar size. This allows us to discard faces which may look similar, but are located at different distances from the camera. The more distant the face, the smaller the size of the bounding box. For each face detected in the left camera, a set of potential matches are found for faces in the right camera.

Both images are template matched against each other, using the first image as a template. The images are scales to the same size and a result is calculated using cross-correlated normalisation method.

$$R(x,y) = \frac{\sum_{x=0,y=0}^{n} T(x,y) \times I(x,y)}{\sqrt{\sum_{x=0,y=0}^{n} T(x,y)^2 \times \sum_{x=0,y=0}^{n} I(x,y)^2}}$$

Unlike the normal usage of Normalised Cross-Correlation, the pixels from the images are only match against the corresponding pixel in the template (i.e. There is no shifting of the pixels in the image to search for a match in the template).

The pair with the highest score is chosen as the face most likely be tracked by the system.

3.5 Should We Track?

Before the system moves into tracking mode, the active head will move to centralise the selected faces. This is to improve the accuracy of the correlation in future passes, and to make the selected user aware that the robot is trying to focus on them. While this centralisation process is taking place, the system is still processing and correlating faces in the scene.

When the selected faces are centralised in both cameras (ie the bounds of the selected faces overlap the central 1600 pixels in each camera, then the system moves on to face tracking.

3.6 Tracking

Although the Viola-Jones detection is efficient enough to allow tracking through sequential detections throughout a sequence, it is sensitive to changes in the pose of the face. As such, we utilise a more efficient approach which is also invariant towards the pose of the face. For this we use the continuously adaptive mean shift (CAMSHIFT) algorithm, a robust non-parametric iterative technique for finding the mode of probability distributions based on the mean shift algorithm [Comaniciu and Meer, 1997]. The method builds an initial colour histogram from the detected face region from which the probability distribution of the face in the colour image can be computed. The algorithm then detects the mode in the probability distribution image closest to its current estimate through the mean shift algorithm whilst dynamically adjusting the parameters of the target distribution. This iterative procedure generally converges within a few iterations.

For our purposes we build the histogram using only the hue of the detected region as the skin hue exhibits good discriminative properties from most other objects encountered in indoor environments. However, to constrain the method in cases where objects surrounding the face exhibit skin-like hue, we restrict the ratio of the area containing the face.

We used the OpenCV implementation of CAMSHIFT here, where the centre of the bounding box was assumed to correspond to the centre of the subjects face. The Active head is then directed to move the cameras so as to minimise the error between the centre of the camera and the centre of the subjects bounding box.

3.7 Re-detection

As the purpose of the system is to interact with the subject while the subject is showing interest in the robot, there needs to be a way of detecting when the subject loses interest in the robot. This loss of interest can take a number of forms, including walking away from the robot (going out of shot), or simply turning away (but remaining in shot). The first situation is simple to deal with as the Camshift algorithm loses track of the subject being tracked and signals that the system should return to the Face detection state. In the second situation the subject remains in shot but turns away from the camera. In this case the Face Tracking software still recognises the subjects skin colour and continues to track the subject, moving the head if necessary. To prevent this occurrence causing the robot to behave in a confused or threaten-

ing manner, the system will periodically attempt to re-detect the subjects face within a restricted area in the frame. The reason for restricting the size of the frame is twofold. Firstly to eliminate the chances of false positives ie The subject face is looking away, but another face has entered the full frame. The second reason is for processing speed, as face detection is a more expensive process than face tracking. The geometry of the frame is specified as

$$F_d = F_t \times G_f$$

with G_f set to allow growth of the tracked image based on the movement of the subject. [2] Within this restricted area the face detection algorithm is run at a rate of F_n per second. [3] If a face is not detected within this frame a counter is incremented to reflect the absence of the face within the frame. If a face is detected within the restricted frame then the counter is reset. If the counter reaches the threshold then the system returns to the face detection state.

3.8 Motor Control Software

The software to control the Active Head motors is built into DROS. The control system consists of a Linux driver for the ServoToGo card, and a PID controller. The PID controller was developed to support position control, and extended by Shaun Press to enable velocity control.

The control of the head is based on minimising the difference between the centre of the detected or tracked face, and the centre of the camera image. The movement velocity of the cameras are calculated by the difference between the centre of the camera centre along the x axis and the image centre along the x axis. Camera movements velocity is calculated as

$$V = \begin{cases} 0, & \text{if } (C_x - I_x) < \epsilon \\ (C_x - I_x) \times D_f, & \text{if } (C_x - I_x) \geq \epsilon \end{cases}$$

where C_x is the centre position of the camera along the x axis, I_x is the centre of the detected face image along the x axis, ϵ is the threshold for movement, and D_f is a dampening factor to reduce the speed of the camera movement to a more "natural" rate. ϵ is set to a small value (between 0 and 0.2 radians) to allow slight movement by the subject without triggering a camera move. This removes the oscillating effect of the cameras

---

[2] Normally G_f is set between the range of 1.0 to 1.2

[3] F_n is usually in the range of 1 and 4 times per second

jittering back and forth due to slight movements of the tracking box.

For tilt control the velocity control formula is

$$V = \begin{cases} 0, & \text{if } (C_y - I_y) < \epsilon \\ (C_y - I_y) \times D_f, & \text{if } (C_y - I_y) \geq \epsilon \end{cases}$$

where $I_y = \frac{I_l y + I_r y}{2}$ ie the mean centre of the two images.

4 Results

To demonstrate the robustness of the system the following experiments were carried out. For all the tests the damping factor for velocity control (D_f) was set to 0.2. This was the setting that the authors believed provided the most "human like" motion of the active head. For higher levels of accuracy a higher value for D_f should be set.

4.1 Face Selection

We tested a couple of important features of the face detection component. Firstly the system has to identify faces very quickly and then correlate them in a short period of time. This is necessary to capture the attention of the user as quickly as possible. If the detection stage takes to long the user may have walked away from the exhibit before the tracking stage can even begin. The results in Figure 4 show the time between recognising a face and when tracking starts. For this experiment there was a single subject who started in front of the camera, but during the course of the experiment moved about. The 'start time' entry (measured in seconds) was when the face detector identified faces in both cameras, and the 'end time' entry was when tracking with the camshift algorithm began. The correlation value was the number of times the face correlation routine was called before tracking began. The final column shows the average length of time to process each call to the correlation function. Of course this time is also taken up with doing other things such as velocity calculations and calls to the face detector, but it does demonstrate how quickly the whole system operates.

Secondly the system must be accurate when matching faces. With multiple faces in the scene both cameras must focus on the same face as otherwise the cameras can end up pointing in different directions, giving the impression that the system is failing to work in a human like manner.

Start Time (s)	End Time (s)	Duration	Correlations	Correlation Speed
0.000	2.414	2.414	11	0.219
5.680	6.810	1.130	8	0.141
10.480	10.677	0.197	1	0.197
13.946	15.128	1.182	10	0.118
22.883	26.014	3.131	5	0.626
29.278	29.480	0.202	1	0.202
34.548	36.011	1.463	10	0.146
39.280	45.880	6.600	24	0.275
55.747	57.481	1.734	11	0.158
60.744	60.943	0.199	1	0.199
70.623	71.475	0.852	6	0.142
75.422	76.060	0.638	1	0.638
79.159	79.355	0.196	1	0.196
83.950	86.274	2.324	17	0.137
90.356	90.546	0.190	1	0.190
94.008	95.875	1.867	16	0.117
101.343	101.741	0.398	1	0.398

Figure 4: Time log of of transition between face detection and face tracking

4.2 Tracking Time

The length of time that a face is tracked is important in the use of the system. The results in Figure 5 show the length of time a single face was tracked by the system. During this test the subject moved about, but stayed within the physical range of the cameras.

The tracking period ran for approximately 100 seconds until the subject walked out of sight of the cameras. In between the times when the system was tracking the subject the system was engaged in either trying to detect the subjects face, or correlate matching faces in both cameras. (See Figure 4

The test was then repeated with two subjects in the scene. Again the focus subjects moved about the scene. This experiment ran for 60 seconds and ended when the subjects moved out of sight of the cameras.

In both experiments the cameras remained focussed on the subject for its entire length. As long as the subject remained in view the cameras wither tracked to subject using Camshift, or actively centred on the subjects face and resumed tracking. In the case of multiple faces in the scene the system quickly determined the correct matching faces and proceeded to track the subject.

Even when the system had not selected a face to track it still moved the cameras in such a way as to indicate that is was trying to look at the subject. Feedback from

Start Time	End Time	Duration
2.414	5.680	3.266
6.810	10.480	3.670
10.677	13.946	3.269
15.128	22.883	7.755
26.014	29.278	3.264
29.480	34.548	5.068
36.011	39.280	3.269
45.880	55.747	9.867
57.481	60.744	3.263
60.943	70.623	9.680
71.475	75.422	3.947
76.060	79.159	3.099
79.355	83.950	4.595
86.274	90.356	4.082
90.546	94.008	3.462
95.875	101.343	5.468

Figure 5: Tracking times using Camshift

Start Time	End Time	Duration
4.275	8.089	3.814
10.543	13.817	3.274
14.789	53.463	38.674

Figure 6: Tracking times using Camshift - with dual faces

the test subjects indicated that they believed that the robot was looking at them, in the case of the chosen face, and they considered the active tracking 'natural'. Another of the test subjects described this behaviour as 'freaky'.

4.3   Tracking Error

The ability of the cameras to follow the user was tested by recoding the error between the required camera position (the centre of the tracked image) and the centre of the camera image. The test subject started off standing in front of the robot at a distance of 2 metres. The subject then moved around in front of the robot, in 3 dimensions (ie left,right, up, down, back and forward). The subject remained in view for approximately 100 seconds before ending the experiment by moving out of view. An indication of a successfully working system was when the error returned to 0 degrees. For operational purposes the movement activation threshold was set to +- 5 degrees (ie the camera or tilt axis would only move if the error between the face centre and the image
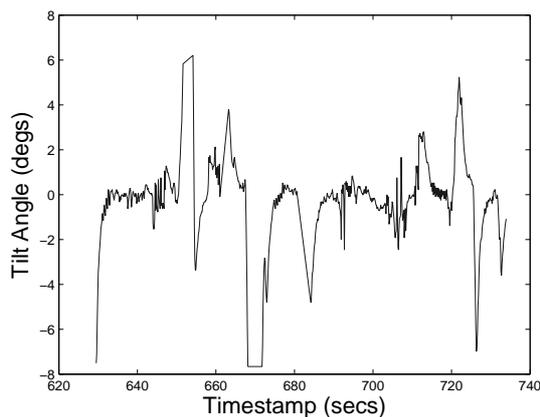
centre exceeded this amount).

All three axis exhibited sensible behaviour in quickly minimising any error between point of focus and centre of face. Each of the axis had their own characteristics that are worth noting. The tilt access (Figure 7) had the best performance but this is a function of the experiment as the height of the subject rarely changes during tracking. The left camera (Figure 8)also had good performance and successfully tracked the more common left-right movement of the subject. The error in the right camera (Figure 9) was more marked but this was caused
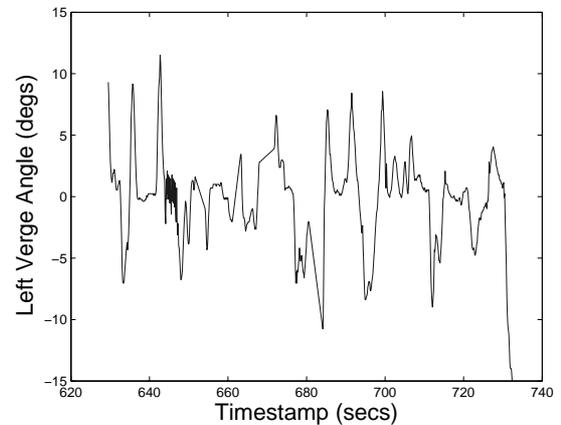

Figure 8: Error (in degrees) between centre of left camera and centre of face

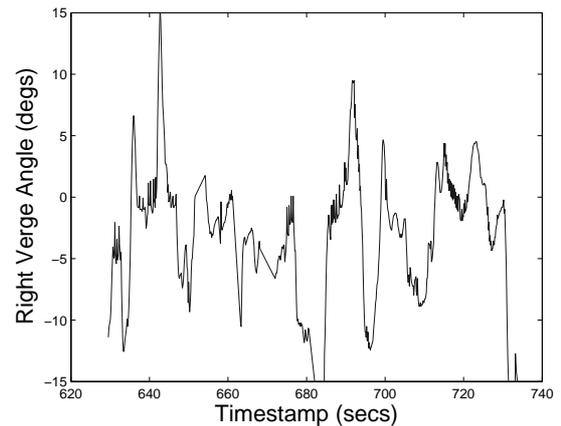Figure 7: Error (in degrees) between Active Head Tilt and centre of face

Figure 9: Error (in degrees) between centre of right camera and centre of face

by a bearing failure in the right axis on the active head, creating slippage in low velocity movement. Nonetheless the right camera was able to successfully track the subject, although with some minor lag.

5   Conclusion

This paper demonstrates a working method of using an active head face tracking system to encourage interaction by humans with robots. The performance of the system is fast enough to convince the user that the Questacon robot is taking an interest in them through the use of eye contact. The system is robust enough to work in a crowded environment, enabling it to be successfully used in public places. With the combination of both eye contact and verbal understanding an response, the Questacon Robotic Exhibit is successful in drawing humans into an interactive experience with the robotic world.

References

[Bradski, 1998] Gary R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, (Q2):15, 1998.

[Comaniciu and Meer, 1997] D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. In *ICCV 97*, pages 750–755, San Juan, Puerto Rico, 1997.

[Petersson *et al.*, 2001] L. Petersson, D. Austin, and H. Christensen. Dca: A distributed control architecture for robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2361–2368, 2001.

[Press *et al.*, 2006] S. Press, J. Chen, and J. Saragih. An interactive speech system for a robotic exhibit. In *Submitted to Australian Conference of Robotics and Automation*, 2006.

[Truong *et al.*, 2000] H. Truong, S. Abdallah, S. Rougeaux, and A. Zelinsky. A novel mechanism for stereo active vision, 2000.

[Viola and Jones, 2001] Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 511–518, 2001.