Vision Systems For Identifying Interlocutor Behaviour And Augmenting Human-Robot Interaction

Pranav Barot — Systems Design Engineering, UW
Ewen MacDonald — Systems Design Engineering, UW
Katja Mombaur — Systems Design Engineering, UW
Email: {pbarot, ewen.macdonald, katja.mombaur}@uwaterloo.ca

Fig. 1: REEM-C Humanoid Robot, with RGB-D camera shown on forehead

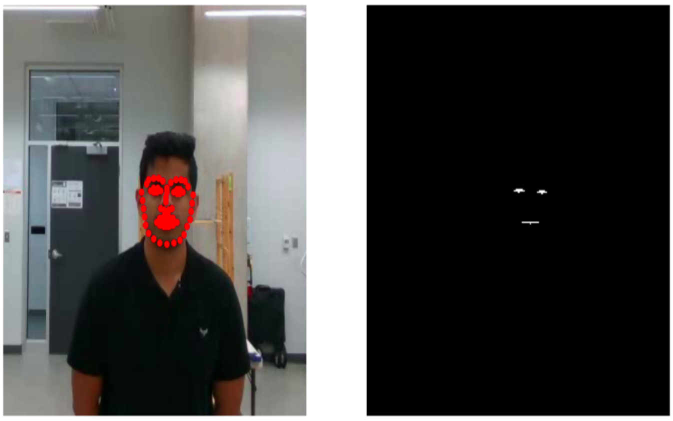

Fig. 2: Detected facial landmarks shown in red (left), and mouth and eye regions shown as binary masks (right)

Abstract

Robots require innovative, intelligent systems to effectively interact with potentially multiple humans simultaneously. A consortium of unique systems may be necessary to properly understand and respond to a variety of human behaviours. This extended abstract presents relevant imaging systems, including visual voice activity detection, gaze estimation, and identification of angular positions of humans relative to the robot. We show that video data alone provides a framework to interact with humans which is of high importance in multi-modal robotics systems. Data collection, processing, and initial results for these algorithms are presented.

1 Introduction

Human-robot interaction systems often involve using either audio, video, or both in tandem to facilitate tasks such as working in assembly workspaces [1] or communicating with human subjects [2]. There is a need to employ these systems for the specific purpose of engaging in conversations with one or more humans present. A robust set of imaging algorithms are therefore required to allow for this functionality.

Voice activity detection is easily accomplished when audio information of present interlocutors is available. The applications of visual voice activity detection involve scenarios when audio is not available due to hardware restrictions, or is unreliable due to noisy scenes or sounds that are recorded simultaneously. Additionally, gaze estimation will be important to identify important conversational cues and find where interlocutors are directing their attention. Alongside a method to estimate the angle of the interlocutor relative to the robot, a framework to allow for a complete human-robot interaction can be developed. These problems are often approached with computationally heavy deep learning models trained on extensive datasets [3, 4], and therefore require a lightweight, classical computer vision alternative.

All systems are designed for the REEM-C Humanoid Robot, which utilizes a RealSense RGB-D camera. Due to bandwidth limitations on the ROS network, these systems are designed for a feed of 15 fps at a resolution of 640 x 480. An image of the REEM-C is shown in Figure 1.

2 Visual Voice Activity Detection

Conversational scenarios with a humanoid robot will require identifying when humans are speaking and when they are silent. Visually, this can be done by identifying features that correlate to speech. Initial data is collected by recording audio and video simultaneously of human conversations from the REEM-C. Audio is recorded in 2 channels, at a standard of 44.1 kHz, and video frames are recorded directly from the REEM-C camera.

To understand which features may correlate with speech, audio data is broken into frames of 1/15th of a second, to match the data extracted from video. Features are directly extracted from frames using facial landmarks from the DLIB detector, which is commonly used in the literature [5]. These landmarks allow for measurement of important characteristics from detected faces, such as those of the mouth or eye areas. Figure 2 shows detected facial landmarks on a subject, and the extracted the mouth and eye regions as binary masks.

Features that may be related to speech are extracted on each frame, which include but are not limited to, mouth height, mouth area, Sobel filter gradients, and HSL information from the mouth areas. A common method to process features includes using a sliding window approach, which moves frame-by-frame, spanning the sequence of collected data. To mimic concepts used in audio signal processing, metrics such as the mean and the power of the window [6] are used to identify behaviour that is of interest. For instance, we suppose that as a person speaks, the area of their mouth will be larger than when they do not speak, which will be reflected in the mouth area signal. The average of the sliding windows is taken for video data, time aligned with audio data, and compared to the audio frame power. For a given window of data, the mean and the power of the window are computed as in Equation 1 and Equation 2. A window size of N=5 frames is chosen for this experiment.

$$Mean(x) = (1/N)\sum_{n=1}^{N} x[n] \tag{1}$$

$$Power(x) = \sum_{n=1}^{N} x^2[n] \tag{2}$$

A correlation matrix is generated to better understand which visual features correlate with two channel audio power. These features are normalized by the subject's mouth size, to accommodate for different facial features between subjects. An example correlation matrix is shown in Figure 3.

Noticeably, certain features have some correlation with audio power, and may be used in tandem to identify periods of voice activity. Lightness pixels indicates the number of pixels in the mouth region below the average lightness of the mouth area detected in the first frame. This feature's extraction is shown in Equation 3, on every frame.

Figure 4 and Figure 5 show window averages for lightness pixels of the mouth region, and the area of the mouth, respectively. Green windows indicate periods of voice activity.

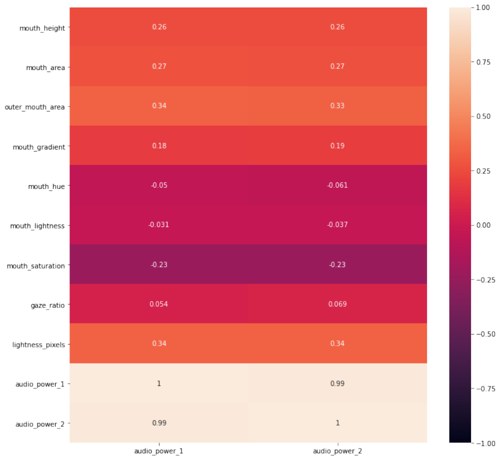

Fig. 3: Sliding window feature means in correlation matrix with audio frame power

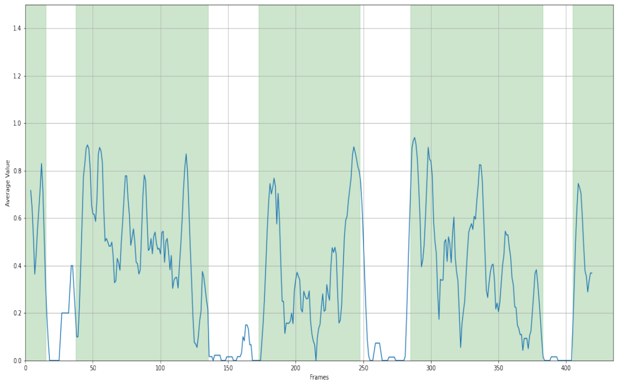

Fig. 4: Sliding window average of lightness pixels feature. Green = voice activity, White = no voice activity

$$
\begin{aligned}
&\text{lightness pixels} = \frac{\sum_{l=1}^{L} P_l < T}{\text{mouth area}} \\
&P_l = \text{l-th pixel of mouth area, in the lightness channel} \\
&T = (1/L)\,(\text{sum of all pixel lightness in mouth of first frame})
\end{aligned} \tag{3}
$$

The threshold T is updated every 10 seconds to account for possible changes in lighting or positioning of the subject. It can be seen that the lightness feature corresponds well with windows of voice activity. A similar pattern is seen when looking at the power of the mouth height measurements. We hypothesize that a fluctuation of values within the periods of voice activity are a result of the pronunciation of different phonemes, requiring the mouth to open and close accordingly.

A measure is used in Equation 4 to capture the increase in values of these features by modeling the data windows as Gaussian distributions [7]. Voice activity is classified as per the following conditions from [7] once the mean and power of each frame is computed. Q is the probability of a Gaussian random variable given the window mean and standard deviation, and PFA (probability of false alarm) is set to 1%.

$$
\begin{aligned}
&Mean(x) > \alpha_1 \ and \ Power(x) > \alpha_2 \\
&\alpha_1 = \sqrt{\frac{\sigma^2}{N}}Q^{-1}(PFA) \\
&\alpha_2 = \sigma^2 Q^{-1}(PFA)
\end{aligned} \tag{4}
$$

This algorithm is applied to a test video with two subjects conversing and generates results with the lightness pixels feature as shown in Figure 6.

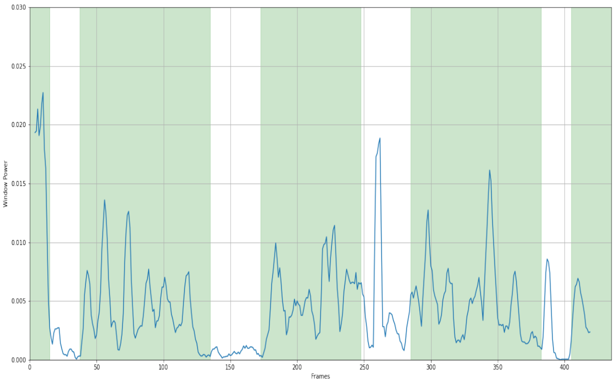

Fig. 5: Sliding window power of mouth heights feature. Green = voice activity, White = no voice activity

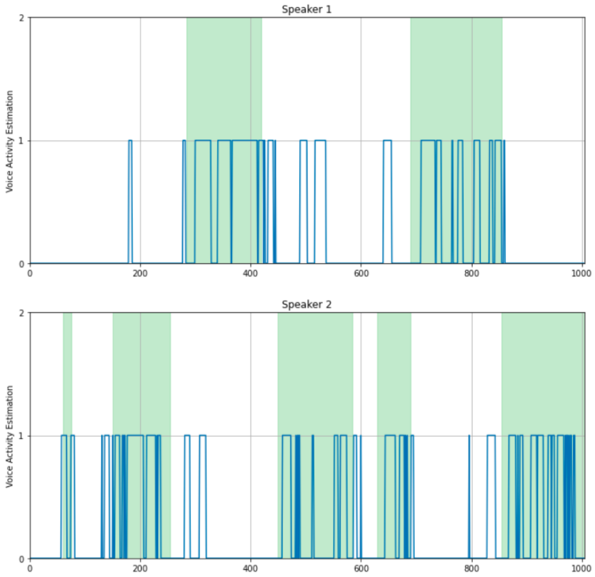

Fig. 6: Voice activity detection estimates, in blue, visualized against annotated windows of speech for speaker 1 and speaker 2 in conversation. Green = voice activity, White = no voice activity

The estimates are dense in periods of speech, indicating the algorithm is able to isolate distinct periods of voice activity for both speakers. The use of other features may be key in reducing false positives and better capturing the entire window of voice activity when a subject speaks.

This algorithm is also functional online, and accommodates for new subjects entering or leaving the field of view of the robot.

3 Gaze Estimation

Gaze is estimated using binary masks on each half of the eye, for both eyes. For the left eye for instance, a binary mask for all pixels enclosed by points (36,37,41) and (38,39,40) are extracted. The amount of sclera present in either half of the eye indicates where the iris may be placed [8], thereby estimating the direction of the subject's gaze. This amount of whiteness in the eye can be measured via the average lightness in the eye halves in the HSL space. The ratio of lightness in the left half of the eye to the lightness in the whole eye is computed, for both eyes. The same is done for the right half of the eyes, and then the two are subtracted to generate a difference in lightness for both halves of the eyes. This is further explained in Equations 5, 6, 7, 8 and 9.

$$
\text{L}_{\text{eye}}, \text{L}_{\text{half}} = \frac{\text{avg. lightness in left half of left eye}}{\text{avg. lightness in left eye}} \tag{5}
$$

$$
\text{R}_{\text{eye}}, \text{L}_{\text{half}} = \frac{\text{avg. lightness in left half of right eye}}{\text{avg. lightness in right eye}} \tag{6}
$$

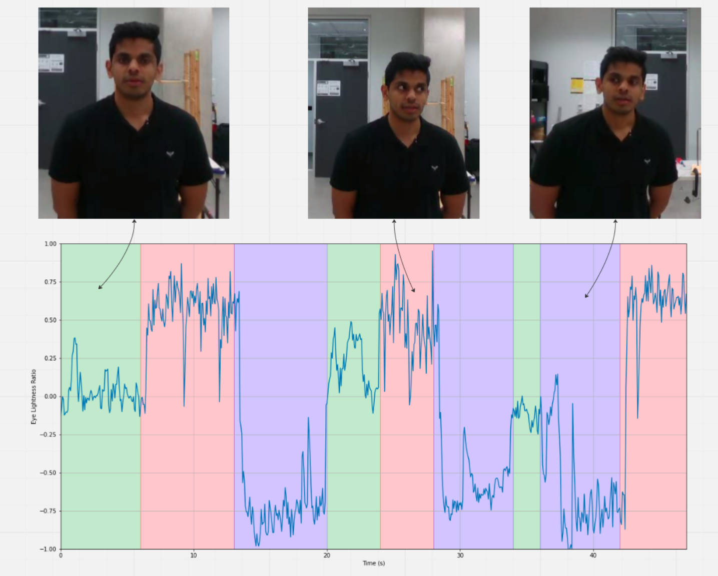

Fig. 7: Gaze ratio plotted against annotated windows of gaze. Red = gaze to the left, blue = gaze to the right, and green = gaze directed forward

$$\mathrm{L_{eye}, R_{half}} = \frac{\text{avg. lightness in right half of left eye}}{\text{avg. lightness in left eye}} \tag{7}$$

$$\mathrm{R_{eye}, R_{half}} = \frac{\text{avg. lightness in right half of right eye}}{\text{avg. lightness in right eye}} \tag{8}$$

$$\begin{aligned}\text{gaze ratio} = & (\mathrm{L_{eye}, L_{half}}) + (\mathrm{R_{eye}, L_{half}}) \\ & - (\mathrm{L_{eye}, R_{half}}) - (\mathrm{R_{eye}, R_{half}})\end{aligned} \tag{9}$$

This gaze ratio is extracted on every frame, and checked against determined thresholds to identify if the person is looking to their right, left, or forward. Figure 7 shows the gaze ratio for a test video, with annotated windows for where the subject's gaze was directed.

From the test data, hard thresholds are imposed to determine the subject's gaze, as shown in Equation 10.

$$\text{gaze direction} = \begin{cases} \text{left} & if \text{ gaze ratio } > 0.5 \\ \text{right} & if \text{ gaze ratio } < -0.5 \\ \text{forward} & else \end{cases} \tag{10}$$

4 Interlocutor Angle Identification

Allowing for realistic conversations also involves identifying where exactly the speakers are, in terms of angular displacement relative to the robot. This can be done by using the RGB-D camera's depth information, combined with the location of the subject's face in the 2D image, to triangulate their position. The location of the subject's face is taken as the average of the coordinates of the landmarks that outline the face. This is demonstrated in Figure 8.

This vision system opens up the possibility of orienting the robot towards the person who is detected to be talking, in a way that replicates natural human behaviour. When head and torso re-orientation is performed, or the robot is made to step, compensation is applied to resolve the relative interlocutor angle to an absolute coordinate system.

5 Discussion

All three systems must work together to facilitate a more complete human robot interaction scenario. Each present subject in the scene is characterized on every frame by their voice activity status, their gaze, and their angle relative to the robot. This information is transmitted through the ROS network, which allows for a framework to be developed for interacting with one or more subjects.

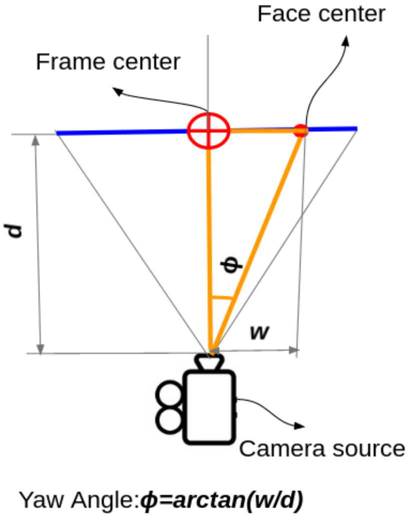

Fig. 8: Identification of interlocutor angle using depth information

There is a large potential for using the extracted information for the purposes of human-robot interaction and conversation [9–11], including gaze following, implementing conversational cues and more.

All of this can be accomplished without audio functionality available, however, to allow for real conversation with humans, audio is desirable. Algorithms for sound source localization [6] can be used to correct for scenarios where faces are undetected, masked, or outside the field of view for the purposes of voice activity detection. Audio also opens up the potential of using speech recognition and chat functionalities to augment the human to robot conversation. Direction of arrival estimates using audio can also be used to correct for errors in visual voice activity detection, by increasing confidence in estimates of who is speaking, or verifying false positives.

6 Conclusion

In conclusion, three unique imaging systems are proposed to better facilitate human-robot interaction. Given good accuracy on these algorithms, they can be synthesized to allow for the REEM-C to intelligently interact with multiple humans simultaneously.

References

[1] A. Bannat, J. Gast, T. Rehrl, W. Rösel, G. Rigoll, and F. Wallhoff, “A multimodal human-robot-interaction scenario: Working together with an industrial robot,” in *Human-Computer Interaction. Novel Interaction Methods and Techniques*, J. A. Jacko, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 303–311.

[2] N. Rastogi, F. Keshtkar, and M. S. Miah, “A multi-modal human robot interaction framework based on cognitive behavior therapy model,” 07 2018.

[3] S. Guy, S. Lathuilière, P. Mesejo, and R. Horaud, “Learning visual voice activity detection with an automatically annotated dataset,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4851–4856.

[4] V. S. Shi L, Copot C, “Gaze gesture recognition by graph convolutional networks,” in *Front Robot AI. 2021 Aug 5*, 2021.

[5] D. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 07 2009.

[6] C. Rascon and I. Meza, “Localization of sound sources in robotics: A review,” *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 08 2017.

[7] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, “Visual lip activity detection and speaker detection using mouth region

intensities," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 133–137, 2009.

[8] C. Wang, Y. Wang, Y. Liu, Z. He, R. He, and Z. Sun, "Sclerasegnet: An attention assisted u-net model for accurate sclera segmentation," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 1, pp. 40–54, 2020.

[9] G. Sandini, A. Sciutti, and F. Rea, *Movement-Based Communication for Humanoid-Human Interaction*, 01 2019, pp. 2169–2197.

[10] K. Lohan, H. Lehmann, C. Dondrup, F. Broz, and H. Kose, *Enriching the Human-Robot Interaction Loop with Natural, Semantic, and Symbolic Gestures*, 09 2017, pp. 1–21.

[11] A. Cangelosi and T. Ogata, *Speech and Language in Humanoid Robots*, 09 2017, pp. 1–32.