Optimization of a Robotics Gaze Control System

Jaime Duque Domingo[1], Jaime Gómez-García-Bermejo[12], and Eduardo Zalama[12]

[1] ITAP-DISA, University of Valladolid, Pl. Santa Cruz 8, 47002 Valladolid, Spain
jaime.duque@uva.es
[2] CARTIF Foundation, Boecillo 47151,
Valladolid, Spain

Abstract. Gaze control is an important issue in the interaction between a robot and humans. In particular, the decision of who to pay attention to in a multi-party conversation is a way of improving a robot's naturalness in human-robot interaction. A system based on a competitive neural network can decide who to look at with a smooth transition in the focus of attention when significant changes in stimuli are produced. One important aspect in this process is the configuration of the different parameters of such a neural network. The weights of the different stimuli have to be computed in order to achieve behavior similar to humans. This article explains how these weights can be obtained by solving an optimization problem. The experiments carried out and some results are also presented.

Keywords: Gaze control, gaze engagement, humanoid robot, ROS, HRI, competitive network, computer vision

1 Introduction

Gaze control represents an important discipline in the development of intelligent social robots so as to achieve higher evaluations of a robot's comprehension and naturalness in human-robot interaction [15]. When robots behave like a person, humans feel more comfortable. An approach to this problem consists in using a competitive neural network that receives different stimuli and returns a stable determination of the focus of attention that must be followed with the robot's eyes. One of the advantages of this approach is that the response of the competitive network is smooth and stable, avoiding erratic behavior. This approach also keeps, in the robot's memory, the information about people who previously interacted with the robot, regardless of whether they have left the robot's field of view. Different factors are taken into account: the human's gaze, who is speaking, pose, proxemics, visual focus of attention, hoarding conversation, habituation, etc. These factors produce stimuli which create a dynamic behavior, giving the human interlocutors the feeling of speaking to another human. However, a problem that emerges is the need to determine the importance of the stimuli so as to be able to decide who should be the focus of attention. We have modeled this

importance through a set of gain weights that are adjusted by optimization from an experimentation data set.

These gain weights represent how each stimulus contributes to a particular output of the neural network. In our experimentation, training is carried out by three people who interact with each other and with the robot. At this step, during an initial training, an external user (teacher) analyzes the sequence and establishes who the robot should pay attention to. This information is written down jointly with recorded stimuli. After the initial training, the gain weights are computed and the robot should pay attention to people as learned in the experiments. To validate the proposed methodology, a robotic head with a projected face that simulates the movement of the eyes has been developed. The result of the competition makes the robot set the focus on the person, using the combined movement of neck and eyes.

The present paper is structured as follows: Section 2 explores the state-of-the-art of the technologies considered in this paper. Section 3 briefly explains how the gaze control works with the competitive network. Section 4 presents the approach for computing the weights of the stimuli. In Section 5, the different experiments and results are reported. Finally, Section 6 notes the conclusions of the presented work and suggests future developments.

2 Overview of the Related Work

Human–Robot Interaction (HRI) is a discipline that allows robots that can communicate and respond to ongoing human communications and behavior to be improved [13]. Gaze control is an important issue concerning HRI. A recent survey of the state of the art in social eye gaze for HRI was presented by [1], distinguishing three different approaches to the problem: Human-focused, centered on understanding the characteristics of human behavior during interactions with robots; Design-focused, which studies how the design of a robot impacts on interactions with humans; and Technology-focused, centered on researching how to build tools to guide the robot's gaze in human interaction. The main challenges in a conversation are the management of attention and turn-taking between partners, controlling the gaze and adopting the right conversational roles [1].

According to [15], when a robot is a listener in a multi-party conversation and tracks the conversation with its gaze, it promotes higher evaluations of its comprehension and naturalness than a robot performing random gazing between speakers. Also, in [3], the authors showed how gaze control more effectively motivates users to repeatedly engage in therapeutic tasks. In addition, as seen in [8], the virtual agents which use turn-taking gaze during conversations are evaluated as more natural and pleasant than others that use none, or a random gaze control in their communication. At the same time, the robots with humanoid features positively influence people's behavior towards the machine and their expectations about its capabilities [24].

There have been different works on gaze control, as in [2], where the authors used a circular array of microphones in a social robot, named Maggie, to determine in which direction the robot should look. An infrared laser was used to obtain the distance with respect to the person so that the robot could move forward/backward. A Robot Assisted Therapy (RAT) method was presented in [20], where a robot perceived different stimuli (visual, auditory and tactile) to track a certain colored object, a face or a directional voice. These two previous works did not address the problem of who to pay attention to in a natural way. Different authors have indicated the benefits of fusing sensory information, such as [20], [26] or [28]. In [26], a person is localized with a robotic head by simultaneously processing visual and audio data. However, these works do not focus on a conversation between multiple participants. In [28], the authors created a system to guide a robot's gaze at multiple humans who were interacting, by adding different stimuli: social features, proxemics values, orientation, and a memory component. This approach considered a limited number of stimuli and the maximum value could change abruptly and lead to erratic changes in the focus of attention. Our proposed method is based on a competitive neural network and fixes this problem, creating smooth transitions between participants.

The stimuli used by a robot can be diverse, but those based on computer vision, audio and memory represent some of the most commonly used. Visual information represents an important aspect of HRI [9]. In particular, the Robot's Field of View (FOV) has to be considered to detect visually people situated in front of the robot's camera. But when people are not situated in front of the robot, some information has to be kept in memory. A hypotheses generation was proposed by [21], inferring the people's position by using peripheral vision. Even though a person was not present in the foveal vision, the robot kept plausible hypotheses about the location. In [25], the authors proposed a dynamic visual memory to store information about objects from a moving camera on board a robot and created an attention system based on where to look to reobserve objects in memory and the need to explore new areas. The Visual Focus of Attention (VFOA) represents who or what people are looking at. As presented in [17], there is a relation between head poses and object locations. Audio represents an important stimulus, since microphone arrays can indicate the direction of the incoming sound. As explained previously, some works using microphone arrays, such as [2], [26], have been used with social robots.

When a robot is interacting with people, their detection is required. Some methods are able to detect human bodies, such as *Haar filters* [27], HoG [5], or *Deep Convolution Neural Networks* (DCNN) [18]. However, in our approach, a face recognition algorithm has been preferred, since the participants are assumed to look directly at the robot or have a slightly turned position to look at other people interacting with it. As indicated for human body detections, Haar classifiers, HoG detectors or Deep Learning based solutions are widely used. Haar classifiers detect faces at different scales, but do not deal with non-frontal faces or occlusions. It also returns a large number of false predictions. The HoG feature descriptor is fast, but does not work with small faces. The DLIB library

[14] implements a CNN face detector using a Maximum-Margin Object Detector [7], which works for different face orientations and occlusions. Recognition can be implemented with Deep Residual Learning algorithms, which are very accurate. DLIB implements a ResNet network with 29 convolution layers and uses a pre-trained model which takes the 68 face landmarks obtained from an image [12].

Moreover, lip activity is important to detect whether a person is speaking or not. When several people are situated in front of the robot and some audio is detected, it is likely that the person who is moving the lips is currently speaking. Some works have explored different techniques to detect lip activity, such as [4], where the degree of disorder of pixel directions around the lips using the optical flow technique is measured; or [22], where a statistical algorithm using two detectors based on noise to characterize visual speech and silence in video sequences is created.

Competitive neural networks [10] can process different inputs and decide the winner in a dynamic and natural way. This kind of network can also have habituation capabilities. However, the problem is how to determine the importance of some stimuli over others. This can be done by a set of weights. These weights have to be configured or learned to replicate human behavior when the same stimuli are produced.

3 Gaze Control with a Competitive Network

Gaze control is usually implemented considering different stimuli ([20], [26], [28]). Our approach for the stimuli integration is the use of an on-center off-surround competitive model [10]. Different stimuli are considered, which are the inputs of the competition at an instant of time t, I_{tkx}, where k is the number of persons who interact and x the number of stimuli. These stimuli have a binary value, indicating whether they are present or not, and are balanced by a weight, w_x. In some stimuli, such as speech detection, the value is activated if the different indicators exceed a threshold.

- I_{tk1} shows if a person k is situated in the robot's FOV at the instant of time t. People situated in front of the robot's FOV are usually candidates to interact with the robot. w_1 is the corresponding weight associated to that stimulus.
- I_{tk2} shows that a person k is considered to be speaking. This stimulus is obtained by performing lip movement detection, based on mouth landmarks. In addition, incoming audio has to be detected in the direction of the person k.
- I_{tk3} shows that a person k is gazing directly at the robot. The pose of a person is used to verify if that person is visually interacting with the robot, as shown in Figure 1. Engagement in an interaction is increased by the *mutual gaze*, a kind of *shared looking* [23].
- I_{tk4} shows that a person k is continuously moving. In a conversation with several people, an individual tends to look at another restless person. This

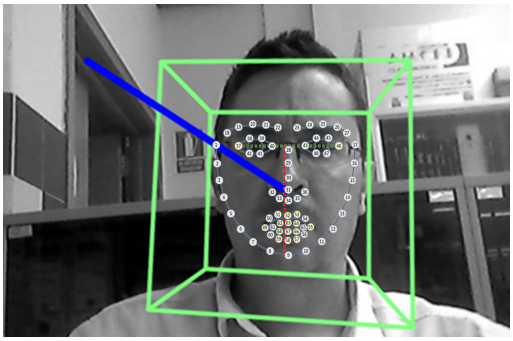

Fig. 1. Pose of a person given the DLIB face landmarks

stimulus requires the individual to be situated in the FOV of the robot. If the sum of the differences in a person's position between several frames is over a given threshold, the person is assumed to be restless.

- I_{tk5} shows that a person k is not situated in the robot's FOV, but for whom incoming audio could have been detected. When the robot does not see a person who has previously interacted, it keeps the previous position of the person in its memory and, if incoming audio is detected in his/her direction, the stimulus for that person k is activated. In a conversation between humans, when someone is speaking at their left/right side, a person tends to turn their head in the direction of the person who is speaking.
- I_{tk6} shows that a person k is in the VFOA of other people, but is not situated in the robot's FOV (see Figure 2). When two or more people are looking at another person in a conversation, a stimulus is given to people in the direction of the gaze. Since the focus of attention is given to a concrete person, the corresponding stimulus is increased.

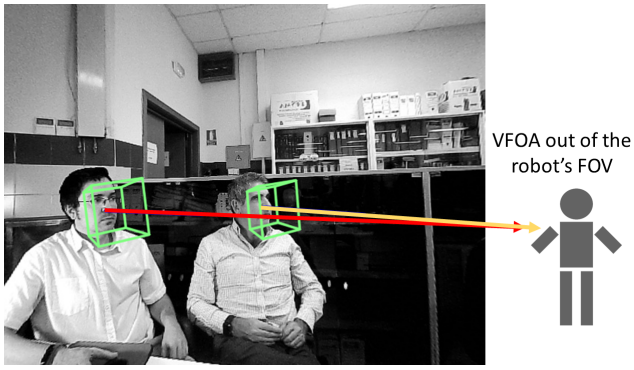

Fig. 2. Two people with their VFOA in another person

– I_{tk7} indicates the proxemics of a person. People situated at a certain distance are likely to be interacting with the robot ([2], [28]). In addition to the weight w_7, this stimulus is balanced by a tuning factor, depending on the distance between the robot and the person.

4 Analysis of the System

The different stimuli that influence a gaze control system must be weighted in such a way that it allows the system to behave similarly to humans. The gaze control system is implemented through a competitive neural network, where each stimulus competes with the others to provide a smooth dynamic result. However, the input of the competitive network, composed by the relation of stimuli and their weights, must be correctly assigned for the robot to have a realistic behavior. Figure 3 shows the system scheme, in which an optimization problem allows the weights that will be used in the gaze control system to be obtained. This optimization requires the data to be split into winners and losers.

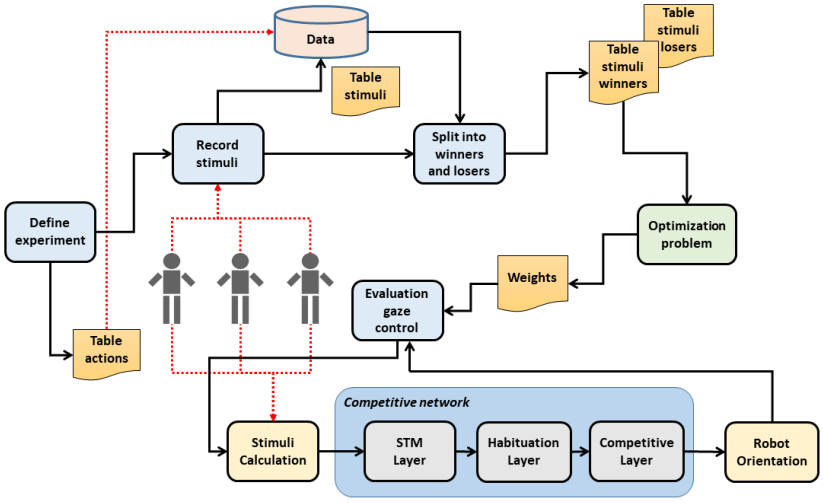

Fig. 3. Scheme of the training system

The competitive network is shown in Figure 4, where each input I_{tk} is obtained by adding the stimuli of the person k, I_{k1},..., I_{k7}, balanced by their respective weights, at time t. Thus, the input of the network produces an output O_{tk} that represents if a person k is the winner. The output with the highest value corresponds to the winner. The configuration of the weights, w_x, is not initially known, so a mechanism has been implemented to obtain it. The competitive

network keeps information about preceding states and dynamically adapts the outputs depending on its configuration.

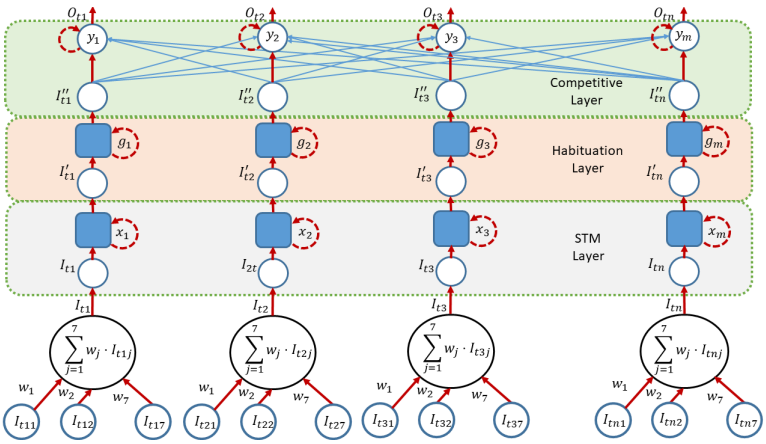

Fig. 4. Competitive neural network

At a given time t, there is an input vector of the network, I_t, composed by the value associated to each person (I_{t1}, ..., I_{tn}). I_{tk} is computed as shown in Equation 1.

$$I_{tk} = \sum_{x=1}^{7} w_x \cdot I_{tkx} \tag{1}$$

The output of the network depends on the input vectors obtained previously, as seen in Equation 2, where C, H and S are, respectively, the layers associated to the competitive, habituation and short time memory (STM) operations.

$$O_t = C(H(S(I_1, ..., I_t))) \tag{2}$$

The STM, habituation and competitive layers are based on the model of Grossberg [10] and use the differential Equations 3, 4 and 5, respectively.

$$\frac{dx_i}{dt} = -A_1 x_i + C_1(B_1 - x_i)[I_i w_i] \tag{3}$$

$$\frac{dg_i}{dt} = E(1 - g_i) - F I_i^{'} g_i \tag{4}$$

$$\frac{dy_i}{dt} = -A_2 y_i + C_2(B_2 - y_i)[I_i^{''} + D{y_i}^2] - y_i \sum_{i \neq j} D{y_i}^2 \tag{5}$$

These equations are solved using a trapezoidal integration, as shown in Equations 6, 7 and 8 for STM, habituation and competition, respectively.

$$\begin{cases} x_i(kh) = x_i((k-1)h) + \frac{g_i(kh)+g_i(k-1)h}{2} \\ g_i(kh) = -A_1 x_i(kh) + C_1 x_i(kh) \\ \qquad\qquad +C_1(B_1 - x_i(kh))[I_i w_i] \end{cases} \tag{6}$$

$$\begin{cases} g_i(kh) = g_i((k-1)h) + \frac{p(kh)+p((k-1)h))h}{2} \\ p(kh) = E[1 - g_i(kh)] - FI_i^{'}(kh) g_i(kh) \end{cases} \tag{7}$$

$$\begin{cases} y_i(kh) = y_i((k-1)h) + \frac{q_i(kh)+q_i(k-1)h}{2} \\ q_i(kh) = -A_2 y_i(kh) + C_2 y_i(kh) \\ \qquad +C_2(B_2 - y_i(kh))[I_i^{''} w_i] \\ \qquad - \sum_{i \neq j} Dy_j((k-1)h)^2 \end{cases} \tag{8}$$

The STM layer, where A_1 corresponds to the decay rate, B_1 to the saturation and C_1 to the growth rate, receives the input stimuli, I_i, and produces an output where the duration of the stimuli is increased. The STM output is the input of the habituation layer, $I_i^{'}$, where the permanent stimuli lose interest through time. In the habituation layer, g_i is the gain for the stimuli. When a stimulus is active, the habituation gain decreases from a maximum value, 1, to the value given by $E/(E + FI_i^{'})$, where E and F correspond to the charge and discharge rates. The output of the habituation layer is the input of the competitive layer, $I_i^{''}$, where A_2 is the decay rate, B_2 the saturation value and C_2 marks the growth rate. D balances the parabolic function ${y_i}^2$, reinforcing the winner against the rest, which represents a lateral inhibition (off-surround). The output of the competitive layer shows the winner of the competition, as explained previously.

The exit from the network takes into account the dynamic nature of previous states by filtering spurious stimuli and producing a smooth change of the winning person focus of attention. However, the input of the network, I_i, corresponding to the stimuli, has to be optimized to obtain a group of weights which are optimal in the gaze control process. To this optimization, the training has to be carried out in a sequential way, following a list of steps previously recorded. The training is carried out by three people. At the same time, an external user (teacher) observes the interaction and annotates the time instants when a person should be the focus of attention.

When all data have been obtained, stimuli from expected winners and losers are separated based on the said manual annotation. At an instant t, there is an input value for the winner, $I_{t,winner}$, and an input value for every k loser, $I_{t,k}$.

The process is modeled as an optimization problem, where the aim is to obtain the optimal weights that maximize the sum of the distances between the winners and the losers at each time instant t, as shown in Equation 9. This procedure ensures that the weights are optimal to make the selected persons winners and separate them from the losers.

$$\begin{aligned} &\max \sum_{t=1}^{m} \left(\sum_{k \in losers} I_{t,winner} - I_{tk} \right) \\ &= \max \sum_{t=1}^{m} \left[\sum_{k \in losers} \left(\sum_{x=1}^{7} w_x \cdot I_{t,winner,x} \right. \right. \\ &\qquad \left. \left. - \sum_{x=1}^{7} w_x \cdot I_{t,k,x} \right) \right] \end{aligned} \tag{9}$$

Some constraints have to be considered. First of all, the sum of all weights has to be equal to 1, as shown in Equation 10. In addition, the weights range between 0 and 1, being bigger than 0 (Equation 11).

$$\sum_{x=1}^{7} w_x = 1 \tag{10}$$

$$\forall x \in [1...7] : w_x \in [0, 1] \wedge w_x > 0 \tag{11}$$

Secondly, there is a constraint for each case evaluated. The input of the winner, $I_{t,winner}$, is bigger than or equal to the losers at an instant t, as shown in Equation 12. With this constraint, the problem is forced to behave as annotated during the training phase.

$$\forall t \in [1..m] \wedge k \in losers \; at \; t : I_{t,winner} \geq I_{t,k} \tag{12}$$

Another consideration is related to the proxemic stimulus. This is balanced by a factor depending on the distance between the person and the robot during the gaze control operation. During the training phase, a fixed value of 1 is assigned. The training is carried out by people who are situated at the *far phase* of the personal space (0.76m to 1.22m) or at the *close phase* of the social space (1.22m to 2.10m) [11], a region where this distance factor is 1. Beyond 2.10 meters, this factor decreases and is not significant for the calculation of weights because, during our training, the participants have been situated within these distances. During normal operation time, an adjustment factor balances the weight of people situated beyond 2.10 meters.

5 Experiments and Results Discussion

The experiment consisted in obtaining the optimal weights during training with three people interacting with the robot. The weights were not initially known and the persons followed a list of actions previously established, as shown in Table 1. The complete list of actions had 726 states. The table shows the losers in green and the winners in red. Only 12 out of the 726 states are shown, but all of them were evaluated in the maximization problem.

Table 1. Sequence of behavior of three people

State /	Person 1							Person 2							Person 3							Winner
Stimuli	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	
1	G						G	R		R				R								2
2	G		G				G	R		R				R	G						G	2
3	R		R	R			R	G		G				G	G		G				G	1
4	G		G				G	R	R	R	R			R	G		G				G	2
5	G		G				G	G		G				G	R			R			R	3
6	R		R	R			R	G		G				G	G						G	1
7	G						G	R	R	R				R	G		G				G	2
8	R		R				R	G						G	G	G					G	1
9	G						G	G		G				G	R						R	3
10	G		G				G	R		R				R								2
11	G						G	G		G				G					R	R		3
12	R	R	R				R	G						G						G		1

The optimization problem was solved using the SLSQP algorithm [16], which obtained the results in 18 iterations and 0.23 seconds (in an intel i9-9900K, with 32Gb of RAM). The obtained results were $w_1 = 0.06$, $w_2 = 0.25$, $w_3 = 0.06$, $w_4 = 0.16$, $w_5 = 0.16$, $w_6 = 0.25$ and $w_7 = 0.06$.

These weights were evaluated in the previously self-developed robotic head (see Figure 5), where the competitive network had been deployed. The network created smooth transitions between the focus of attention, resulting in a natural behavior but, at the same time, considering properly which stimuli were more important according to the obtained weights. Figures 6 and 7 show the sum of input stimuli balanced by their weights, I_{tk}, and the output stimuli, O_{tk}, respectively.

As can be seen at the output of the neural network, the input stimuli are smoothed out to avoid sudden changes in the robot's focus. The results of the winning person are consistent with the choice made during the learning phase, being satisfactory in 99.03% of the test cases. The adjustment of these weights has allowed the creation of a robot that responds in a similar way to how a person behaves.

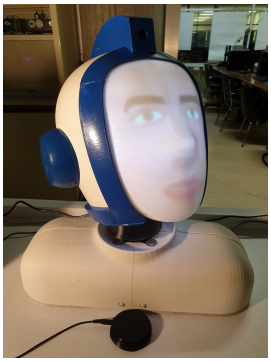

Fig. 5. Self-developed robot head

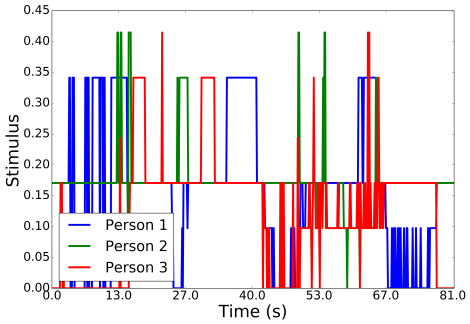

Fig. 6. Input stimuli of three people balanced by the weights

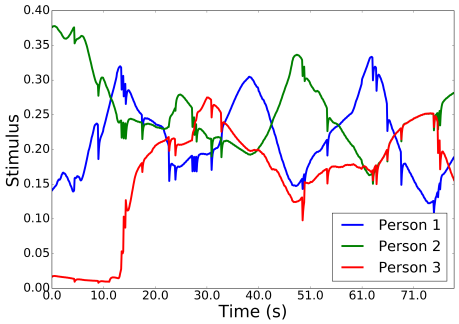

Fig. 7. Output stimuli of three people after the competitive network

Although the system has been trained with 3 people, the stimuli generated during the training are independent of the person. The methods of obtaining stimuli that have been used are generic and suitable for any person. The 726 cases used allow to refine the weights to a very wide range of situations. Training with more people may slightly vary these weights but not in a significant way. At normal operation time, the system works with 3 or more people, having even performed some test with more than 8 participants. Beyond the increase in computing time, the system was able to correctly gaze at the participants who were winning the competition. For performance reasons, our system has been limited to 10 participants.

When a participant disappears from the robot's FOV, the head is able to estimate the position using a Kalman filter [19]. If the robot interacts with 3 participants and one disappears, either because the person moves or because the head turns, when audio is detected in his direction or the other participants are looking at him (VFOA), stimuli will be provided to that participant and the head will gaze at this person in case of being winner, using the estimated Kalman position. When no stimuli are received from a participant for more than 20 seconds, the participant is removed. When a new person arrives on the scene, regardless of whether the robot is able to detect him in the FOV or not, if audio is detected in his direction or if this person is in the VFOA of other participants, the new participant will enter the competition. If this new person becomes a winner, the head will turn to look for him. If no one is detected, it will go back to the old winner.

Although the main aim of this work has been to obtain the weights of our neural network, the head has quantitatively behaved as we expected in operation time in 99.03% of cases, higher compared to other works, such as [28], where the authors obtained results close to 89.4% but with different experiment conditions. A complete explanation about the functioning of the neural network and the robotic head as well as the experiments performed and results obtained are presented in [6].

6 Conclusions

This work presents an optimization process for a robotic gaze control system. This gaze control system uses a competitive network which receives a large number of visual, auditory or presence stimuli. It allows a smooth transition, changing the focus of attention between participants, avoiding erratic movements. In addition, it has habituation capabilities to avoid someone from hoarding the conversation. The weights of each stimulus are not known a priori and a strategy based on an optimization problem has been developed to obtain them through experiential learning.

The computed weights were integrated into the gaze control system of a self-developed robotic head with combined body and eye movement, and the robot showed a natural interaction behavior similar to those annotated during the learning phase. The proposed method significantly improved the sensation of

naturalness and realism of the robotic head. The movement of the robot joints and the expressions of the virtual agent projected on its 3D facial display were controlled by the system integrated in a ROS-based architecture. The design of the robot and the gaze control system creates a more realistic HRI system, which is more acceptable to interlocutors than other not-so-human robots.

The future objectives of the project will consist in integrating speech capabilities into the ROS architecture, offering a low-cost, intelligent robot with human-like behavior.

References

1. Admoni, H., Scassellati, B.: Social eye gaze in human-robot interaction: a review. Journal of Human-Robot Interaction 6(1), 25–63 (2017)
2. Alonso-Martín, F., Gorostiza, J.F., Malfaz, M., Salichs, M.A.: User localization during human-robot interaction. Sensors 12(7), 9913–9935 (2012)
3. Andrist, S., Mutlu, B., Tapus, A.: Look like me: matching robot personality via gaze to increase motivation. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems. pp. 3603–3612. ACM (2015)
4. Bendris, M., Charlet, D., Chollet, G.: Lip activity detection for talking faces classification in tv-content. In: International Conference on Machine Vision. pp. 187–190 (2010)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection (2005)
6. Duque-Domingo, J., Gómez-García-Bermejo, J., Zalama, E.: Gaze control of a robotic head for realistic interaction with humans. Frontiers in Neurorobotics 14, 34 (2020)
7. E. King, D.: Max-margin object detection. arXiv preprint arXiv:1502.00046 (2015)
8. Garau, M., Slater, M., Bee, S., Sasse, M.A.: The impact of eye gaze on communication using humanoid avatars. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 309–316. ACM (2001)
9. Gergle, D., Kraut, R.E., Fussell, S.R.: Using visual information for grounding and awareness in collaborative tasks. Human–Computer Interaction 28(1), 1–39 (2013)
10. Grossberg, S.: Contour enhancement, short term memory, and constancies in reverberating neural networks. In: Studies of mind and brain, pp. 332–378. Springer (1982)
11. Hall, E.T., Birdwhistell, R.L., Bock, B., Bohannan, P., Diebold Jr, A.R., Durbin, M., Edmonson, M.S., Fischer, J., Hymes, D., Kimball, S.T., et al.: Proxemics [and comments and replies]. Current anthropology 9(2/3), 83–108 (1968)
12. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1867–1874 (2014)
13. Kiesler, S., Hinds, P.: Introduction to this special issue on human-robot interaction. Human–Computer Interaction 19(1-2), 1–8 (2004)
14. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research 10(Jul), 1755–1758 (2009)
15. Kousidis, S., Schlangen, D.: The power of a glance: Evaluating embodiment and turn-tracking strategies of an active robotic overhearer. In: 2015 AAAI Spring Symposium Series (2015)
16. Kraft, D., Schnepper, K.: Slsqpa nonlinear programming method with quadratic programming subproblems. DLR, Oberpfaffenhofen (1989)

17. Massé, B.: Gaze direction in the context of social human-robot interaction. Ph.D. thesis (2018)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
19. Rosales, R., Sclaroff, S.: Improved tracking of multiple humans with trajectory prediction and occlusion modeling. Tech. rep., Boston University Computer Science Department (1998)
20. Saldien, J., Vanderborght, B., Goris, K., Van Damme, M., Lefeber, D.: A motion system for social and animated robots. International Journal of Advanced Robotic Systems 11(5), 72 (2014)
21. Shiomi, M., Kanda, T., Miralles, N., Miyashita, T., Fasel, I., Movellan, J., Ishiguro, H.: Face-to-face interactive humanoid robot. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566). vol. 2, pp. 1340–1346. IEEE (2004)
22. Siatras, S., Nikolaidis, N., Krinidis, M., Pitas, I.: Visual lip activity detection and speaker detection using mouth region intensities. IEEE Transactions on Circuits and Systems for Video Technology 19(1), 133–137 (2008)
23. Sidner, C.L., Kidd, C.D., Lee, C., Lesh, N.: Where to look: a study of human-robot engagement. In: Proceedings of the 9th international conference on Intelligent user interfaces. pp. 78–84. ACM (2004)
24. Thrun, S.: Toward a framework for human-robot interaction. Human–Computer Interaction 19(1), 9–24 (2004)
25. Vega, J., Perdices, E., Cañas, J.: Robot evolutionary localization based on attentive visual short-term memory. Sensors 13(1), 1268–1299 (2013)
26. Viciana-Abad, R., Marfil, R., Perez-Lorenzo, J., Bandera, J., Romero-Garces, A., Reche-Lopez, P.: Audio-visual perception system for a humanoid robotic head. Sensors 14(6), 9522–9545 (2014)
27. Viola, P., Jones, M., et al.: Rapid object detection using a boosted cascade of simple features. CVPR (1) 1, 511–518 (2001)
28. Zaraki, A., Mazzei, D., Giuliani, M., De Rossi, D.: Designing and evaluating a social gaze-control system for a humanoid robot. IEEE Transactions on Human-Machine Systems 44(2), 157–168 (2014)