Multimodal attention system for an interactive robot

O. Déniz, M. Castrillón, J. Lorenzo, M. Hernández, J. Méndez

Instituto Universitario de Sistemas Inteligentes (IUSIANI)
Univ. Las Palmas de Gran Canaria
Edif. Central del Parque Científico-Tecnológico. Campus de Tafira
35017 Las Palmas - Spain
odeniz@dis.ulpgc.es

Abstract Social robots are receiving much interest in the robotics community. The most important goal for such robots lies in their interaction capabilities. An attention system is crucial, both as a filter to center the robot's perceptual resources and as a mean of letting the observer know that the robot has intentionality. In this paper a simple but flexible and functional attentional model is described. The model, which has been implemented in an interactive robot currently under development, fuses both visual and auditive information extracted from the robot's environment, and can incorporate knowledge-based influences on attention.

1 Introduction

In the last years the robotics community has sought to endow robots with social and interaction abilities, with the first survey recently published [6]. Researchers realized that robots that excelled in certain tasks were by no means considered intelligent by the general public. Social abilities are now considered very important in order to make the robots more human. Emotion and multimodal communication are also two related aspects that are still being researched.

In [11] the authors argue that a robot with attention would have a minimal level of intentionality, since the attentional capacity involves a first level of goal representations. Attention is a selection process whereby only a small part of the huge amount of sensory information reaches higher processing centers. Attention allows to divide the visual understanding problem into a rapid succession of local, computationally less expensive, analysis problems. Human attention is divided in the literature into two functionally independent stages: a preattentive stage, which operates in parallel over the whole visual field, and an attentive stage, of limited capacity, which only processes an item at a time. The preattentive stage detects intrinsically salient stimuli, while the attentive stage carries out a more detailed and costly process with each detected stimulus. The saliency values of the attentive stage depend on the current task, acquired knowledge, etc [8,10].

Probably the first robot that was explicitly designed to include some social abilities is Kismet [1]. Kismet has had undeniable success in the robotics community because it has been a serious effort in making a robot sociable. Among other diverse modules, Kismet included an attention system, which is based on Wolfe's "Guided Search 2.0

(GS2)" model [15]. GS2 is based on extracting basic features (color, motion, etc.) that are linearly combined in a saliency map. In a winner-take-it-all approach, the region of maximum activity is extracted from the saliency map. The focus of attention (FOA) will then be directed to that region.

It is a well accepted fact that attention is controlled both by sensory salient and cognitive factors (knowledge, current task) [2]. The effect of the lower level subsystem (bottom-up influence) has been comprehensively studied and modelled. In contrast, the effect of higher level subsystems (top-down influence) in attention is not yet clear [9]. Hewett [8] also suggests that volitive processes should control the whole attention process, even though some of the controlled mechanisms are automatic in the human brain. Therefore, high-level modules should have total access to the saliency map. This would allow the attention focus to be directed by the point that a person is looking at, deictic gestures, etc. Fixations to the point that a person is looking at are useful for joint attention. In [14] an additional feature map is used for the purpose of assigning more saliency to zones of joint attention between the robot and a person.

In the third version of Wolfe's Guided Search [16] high-level modules act in two ways. On the one hand they can modify the combination weights. On the other hand, they can also act after each fixation, processing (recognizing, for example) the area of the FOA, after which an "inhibition of return" (IR) signal is generated. IR is a signal that inhibits the current FOA, so that it will not win in the saliency map for some time.

Top-down influences on attention are also accounted for in the FeatureGate model [5]. In this model, a function is used to produce a distance between the low-level observed features and those of the interest objects. In [13] the top-down influence is embedded in the changing parameters that control a relaxation and energy minimization process that produces the saliency map. Also, in [3] a neural network, controlled by high-level processes, is used to regulate the flow of information of the feature maps towards the saliency map. A model of attention similar to that of Kismet is introduced in [12] for controlling a stereo head. Besides the feature maps combination (color, skin tone, motion and disparity), space variant vision is used to simulate the human fovea. However, the system does not account for top-down influences. Moreover, it uses 9 Pentium processors, which is rather costly if the attention system is to be part of a complete robot.

In [7] an attention system is presented where high-level modules do influence (can act on) the whole saliency map. When, after a fixation, part of an object is detected, saliency is increased in other locations of the visual field where other parts of the object should be, considering also scaling and rotation. This would not be very useful in poorly structured and dynamic environments. In the same system, a suppression model equivalent to IR is used: after a fixation the saliency of the activated zone is decreased in a fixed amount, automatically.

The objective of this work was not to achieve a biologically faithful model, but to implement a functional model of attention for a social robot. This paper is organized as follows. Section 2 describes the proposed attention system, implemented for a social robot that is currently being developed. Experiments are described and analyzed in Section 3. Finally, the main conclusions are summarized in Section 4.

2 Attention model

In all the citations made above, the effect of high-level modules is limited to a selection or guiding of the bottom-up influence (i.e. combination weights) and the modification of the relevance of the object in the FOA. We propose that the influence of high-level modules on attention should be more direct and flexible. Inhibition should be controlled by these modules, instead of being an automatic mechanism. The following situation is an example of such case: if I look at a particular person and I like her, inhibition should be low, in order to revisit her soon. There could even be no inhibition, which would mean that I would keep on looking at her. Note that by letting other processes control the saliency map joint attention and inhibition of return can be implemented. Also, the mechanism explained before that increases saliency in the zones where other parts of objects should be can be implemented. In fact, any knowledge-directed influence on attention can be included.

The objective of this work was to conceive a functional attention mechanism that includes sound and vision cues. Therefore, the model proposed here is simple to implement, being the most complex calculations done in the feature extraction algorithms. The activation (i.e. saliency) values are controlled by the following equation:

$$A(p,t) = \sum_i F_i(v_i \cdot f_i(p,t)) + \sum_j G_j(s_j \cdot g_j(p,t)) + K \cdot C(p,t) + T(p,t) \qquad (1)$$

where F and G are functions that are applied to the vision-based (f_i) and sound-based (g_j) feature maps in order to group activity zones and/or to account for the error in the position of the detected activity zones. Spatial and temporal positions in the maps are represented by the p and t variables. v_i, s_j and K are constants. C is a function that gives more saliency to zones near the current FOA: $C(p,t) = e^{-\gamma|p-FOA(t-1)|}$. $T(p,t)$ represents the effect of high-level modules, which can act over the whole attention field. The maximum of the activation map defines the FOA, as long as it is larger than a threshold U:

$$FOA(t) = \begin{cases} \max_p A(p,t) & if \max_p A(p,t) > U \\ FOA(t-1) & otherwise \end{cases} \qquad (2)$$

The model is depicted in Figure 1, using sound and vision for extracting feature maps. Note that a joint attention mechanism would use the component T of Equation 1, which for all practical purposes is equivalent to the approach taken in [14] that used a feature map for that end.

The implementation presented in this paper will use an auditive feature map: the localization of a single sound source. Notwithstanding, this scheme can be used with multiple sources, as long as they are separated by another technique.

The visual feature map is extracted from images taken with an omnidirectional camera, using adaptive background differences. The aim was to detect blobs pertaining to people around the robot. The first step is to discard part of the captured image, as we want to watch only the frontal zone, covering 180 degrees from side to side (see Fig.

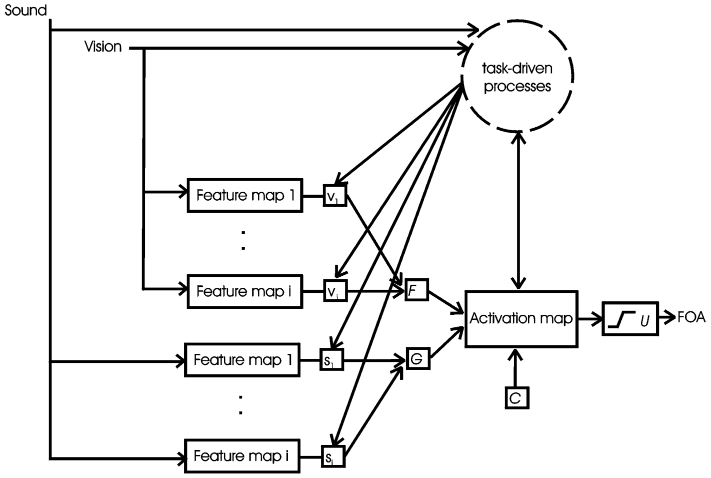

Figure 1. Model of attention. The feature maps must represent the same physical space than the activation map. If sensors do not provide such values, a mapping would have to be done.

2). The background model is obtained as the mean value of a number of frames taken when no person is present in the room. The model M is updated with each input frame:

$$M(k+1) = M(k) + U(k) \cdot [I(k) - M(k)], \tag{3}$$

where I is the input frame. U is the updating function:

$$U(k) = exp(-\beta \cdot D(k)), \tag{4}$$

with:

$$D(k) = \alpha \cdot D(k-1) + (1-\alpha) \cdot |I(k) - I(k-1)|, \tag{5}$$

for α between 0 and 1. The parameters α and β control the adaptation rate.

The method of adaptive background differences described above still had a drawback. Inanimate objects should be considered background as soon as possible. However, as we are working at a pixel level, if we set the α and β parameters too low we run the risk of considering static parts of animate objects as background too. This problem can be alleviated by processing the image D. For each foreground blob, its values in D are examined. The maximum value is found, and all the blob values in D are set to that level. With this procedure the blob only enters the background model when all its pixels remain static. The blob does not enter the background model if at least one of its pixels has been changing.

As for the sound-based feature map, the aim was to detect the direction of sound sources (i.e. people). The signals gathered by a pair of microphones are amplified and preprocessed to remove noise. Then the angle in the horizontal of a sound source is extracted using the expression:

$$angle = \arcsin((s \cdot I/f)/d), \tag{6}$$

where s is the sound speed, f is the sampling frequency, d is the distance between the pair of microphones, and I is the interaural time difference (ITD). The ITD is a measure of the displacement between the signal gathered at one microphone and the signal gathered at the other, and is obtained through correlation. The implemented sound localization system is described in more detail in [4].

3 Implementation and Experiments

The attention model has been implemented on the robot head shown in Figure 2. This head includes an omnidirectional camera as a presence detector and a sound localization system based on a pair of microphones placed on both sides of the head. The feature and activation maps represent a half-plane in front of the robot. The FOA is used to command the pan and tilt motors of the robot's neck. For our particular implementation we decided that sound events should not change the FOA on their own, but they should make the nearest visual event win. Also, as a design decision we imposed that the effect of sound events should have precedence over the effect of C.

In our particular case the variable p takes values in the range $[0,180]$ degrees and F will not be used. $v_1 = 1, f_1 = \{0,1\}$ represents the effect of a visual feature map that detects foreground blobs using adaptive background differences and the omnidirectional camera. The visual feature maps are not actually 1-D, but 1 1/2-D, as for each angle we store the height of the blob, measured by the omnidirectional vision system. This height is used to move the tilt motor of the robot's neck. $g_1 = \{0,1\}$ represents the output of the sound localization routine. The vision and sound localization modules communicate with the attention module through TCP/IP sockets. To account for errors in sound localization, G is a convolution with a function $e^{(-D \cdot |x|)}$, D being a constant. In order to meet these conditions the following should be verified:

- $s_1 < 1$ (the FOA will not be directly set by the sound event).
- Suppose that 2 blobs are anywhere in the activation map. Then a sound event is heard. One of the blobs will be closer to the sound source than the other. In order to enforce the preferences mentioned above, the maximum activation that the farthest blob could have should be less than the minimum activation that the nearest blob could have. This can be put as $1 + K + s_1 \cdot e^{(-D \cdot a)} < 1 + K \cdot e^{(-180 * \gamma)} + s_1 \cdot e^{(-D \cdot b)}$, b and a being the distances from the blobs to the sound source, the largest and the shortest one, respectively. That equation does not hold for $b < a$ but it can be verified for $b < a - \varepsilon$, with a very small ε.

Operating with these two equations the following valid set of values was obtained: $D = 0.01, K = 0.001, s_1 = 0.9, \gamma = 0.15$. For those values $\varepsilon = 0.67$ degrees, which we considered acceptable.

The effect of high-level processes (T) is not used in the implementation yet, as the robot is still under development. The simplicity of the model and of the implementation make the attention system efficient. With maps of 181 values, the average update time for the activation map was 0.27ms (P-IV 1.4Ghz). In order to show how the model performs, two foreground objects (a person and a coat stand) were placed near the robot. A sample image taken by the omnidirectional camera are shown in Figure 2. Initially, the FOA was at the coat stand. Then the person makes a noise and the FOA shifts, and remains fixating the person. In order to see what happens at every moment this situation can be divided into three stages: before the sound event, during the sound event and after the sound event.

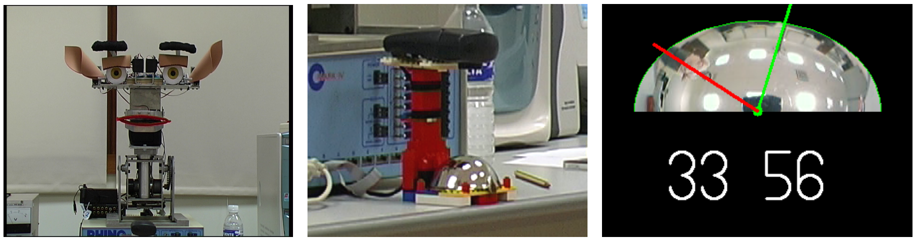

Figure 2. Left: the interactive robot being developed. Center: omnidirectional camera, placed in front of the robot. Right: image taken by the omnidirectional vision system. The numbers indicate the estimated height and the angle of the closest blob (the one with the largest height).

Figure 3 shows the state of the feature maps and the activation map at each stage. Note that the vertical axis is shown in logarithmic coordinates, so that the effect of the *C* component, which is very small, can be seen. The exponential contributions thus appear in the figures as lines.

Before the sound event the FOA was at the blob on the left, approximately at 75 degrees, because it is the closest blob to the previous FOA (the robot starts working looking at his front, 90 degrees). This is shown in the first two figures. The two next figures show the effect of the sound event. The noise produces a peak near the blob on the right (the person). That makes activation rise near that blob, which in turn makes the blob win the FOA. The last two figures show how the FOA has been fixated to the person. In absence of other contributions the effect of the *C* component implements a tracking of the fixated object/person.

4 Conclusions

An attentional system is a necessary module in a complex human-like robot. With it, the robot will be able to direct its attention to people in the environment, which is crucial for interaction. In this paper a simple yet functional model of attention has been described, drawing upon previous attentional systems for interactive robots. The model was implemented using both auditive and visual features extracted from a zone surrounding the robot. Visual features were extracted from video taken with an omnidirectional camera, which gives the robot a 180 degrees attentional span. The attentional system is currently running on a robotic head

The next step in our work will be to implement the high-level influences on the attention focus. This influence is to be defined by the robot's tasks and knowledge, which obviously need the completion of other modules, such as an action selection mechanism (with goals), memory and facial analysis.

Acknowledgments

The first author is supported by grant *D260/54066308-R* of Universidad de Las Palmas de Gran Canaria. This work was partially funded by research projects *PI2000/042* of Gobierno de Canarias and *UNI2002/16* of Universidad de Las Palmas de Gran Canaria.

References

1. Cynthia L. Breazeal. *Designing social robots*. MIT press, Cambridge, MA, 2002.
2. M. Corbetta and G.L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3:201–215, March 2002.
3. P. Van de Laar, T. Heskes, and S. Gielen. Taks-dependent learning of attention. *Neural networks*, 10(6):981–992, 1997.
4. O. Deniz, J. Cabrera, and M. Hernández. Building a sound localization system for a robot head. *Revista Iberoamericana de Inteligencia Artificial*. To appear.
5. J.A. Driscoll, R.A. Peters II, and K.R. Cave. A visual attention network for a humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robotic Systems*, Victoria, B.C., Canada, October 1998.
6. T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3-4), March 2003.
7. T.D. Grove and R.B. Fisher. Attention in iconic object matching. In *Procs. of British Machine Vision Conference*, pages 293–302, Edinburgh, September 1996.
8. D. Heinke and G.W. Humphreys. Computational models of visual selective attention: A review. In G. Houghton, editor, *Connectionist models in psychology*.
9. L. Itti. Modeling primate visual attention. In J. Feng, editor, *Computational Neuroscience: A Comprehensive Approach*, Boca Ratón. CRC Press.
10. L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, March 2001.
11. L. Kopp and P. Gärdenfors. *Attention as a Minimal Criterion of Intentionality in Robots*, volume 89 of *Lund University Cognitive Studies*. 2001.
12. G. Metta. An attentional system for a humanoid robot exploiting space variant vision. In *IEEE-RAS International Conference on Humanoid Robots*, pages 359–366, Tokyo, November 2001.
13. R. Milanese, H. Wechsler, S. Gil, J.M. Bost, and T. Pun. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Procs. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 781–785, 1994.
14. B. Scassellati. *Foundations for a Theory of Mind for a Humanoid Robot*. PhD thesis, MIT Department of Computer Science and Electrical Engineering, May 2001.
15. J.M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1(2):202–238, 1994.
16. J.M. Wolfe and G. Gancarz. *"Guided Search 3.0"*, pages 189–192. Basic and Clinical Applications of Vision Science. Kluwer Academic, Netherlands, 1996.

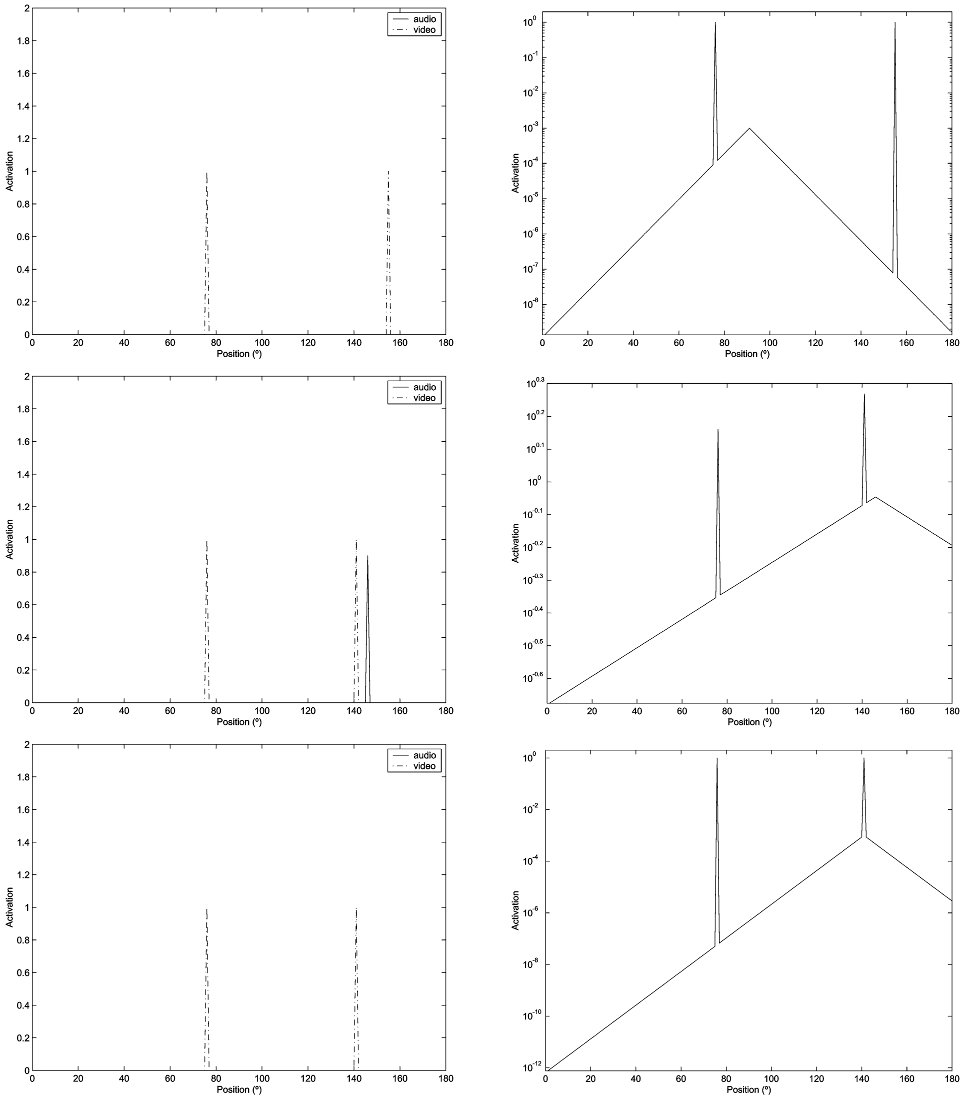

Figure 3. State of the feature and activation maps. On the left column the figures show the visual and auditive feature maps. On the right column the figures show the resultant saliency map.