ICSC 2006
Dynamics in Spatial Interactions

ORAL PAPER

A biologically plausible robot attention model, based on space and time

Anna Belardinelli • Fiona Pirri

Abstract In this work we describe a biological inspired approach to robot attention, developed on the basis of experiments aimed to map human visual search onto robot behaviour, allowing particularly for depth as a further feature in the attention model. By means of a purposely-designed machine we studied fixation zones elicited from scanning paths that were performed during a task driven wandering of subjects' gaze over a cluttered scene. Hence, we defined preference criteria and a utility function accounting for the optimization of visual endeavours. This function would allow a robot to select meaningful spots without the need to process the whole scene.

Keywords Human–robot interaction · Vision-based attention · Spatial cognition

Introduction

Visual attention is a crucial skill for robots and computer vision systems to be endowed with, as it allows an optimal deployment of visual and processing resources only on interesting parts of the visual field.

Visual attention has been deeply investigated to assess how human beings orient gaze and which strategies are applied in order to focus rapidly on salient regions. This ability is crucial in several automatic tasks too, such as search and surveillance, besides object recognition and human–robot interaction.

Treisman and Gelade (1980) proposed a model that gained most credit in past years. According to their framework mechanisms of focused attention are driven by several features such as colour, intensity, and edge orientations, separately or conjunctly perceived. Inspired by this insight, Itti and Koch (2001), Niebur et al. (2001), among others, modelled features extraction and recombination in artificial systems leading to the construction of saliency maps, as a bottom-up control tool for attention in autonomous agents.

Spatial attention drives gaze orienting, particularly in visual search tasks, where a top-down component is present, in that the subject knows approximately what he is looking for and what he expects to see. Attention proceeds along a sequence of fixations and gaze shifts, while observing a scene. Indeed, the fovea can rapidly move from an object to another nearby that gained interest according to its feature properties. Throughout these movements visual information is not processed, being rather left out of the focusing process, a phenomenon known as *change blindness* (Simons and Levin 1997; Rensink et al. 2000). This mechanism allows the human brain to build a sampled, discrete representation of the space instead of a continuous, detailed one. Analogously, in artificial systems, frames collected by cameras on robot heads performing fast rotational and translational movements can be filtered out of the attention process, as they will be blurred and not meaningful, saving in this way visual and processing resources. In our experiments we studied fixations occurred in a task-driven wandering of the gaze over wide scenes. Since these were 3D environments, displaying a large set of depth

A. Belardinelli (✉) · F. Pirri
Dipartimento di Informatica e Sistemistica,
Universita' di Roma "La Sapienza", Rome, Italy
e-mail: belardinelli@dis.uniroma1.it

F. Pirri
e-mail: pirri@dis.uniroma1.it

planes we considered how depth would intervene in scanning strategies.

The way depth affects attention deployment has been researched only in recent years in the field of Cognitive Sciences. In this sense, Theeuwes et al. (1998) showed that mechanisms grouping contiguous objects work discriminating different depth planes. Several authors (Previc 1998; Maringelli et al. 2001; Couyoumdjian et al. 2003) suggested that different mental representations are used for far and near space according to the diverse specificity of the performable tasks (mainly perceptual or motor, respectively).

In artificial systems, depth was integrated as an ulterior feature in the construction of the saliency map by Ouerhani and Hügli (2000), while Frintrop et al. (2004) used visual attention onto range images within an object recognition task.

The above considerations underpin our approach to strategy based task-driven attention to attain a model for the automatic generation of likely scanning paths. Paths are defined in terms of a sequence of rotation angles to be performed by the robot head equipped with a stereo camera. Fixation zones can be determined by means of a utility function deduced once the data recorded from subjects' performances (fixation points and gaze velocities) were reported on a mosaic scene.

Experimental setup and data processing

To record data from a subject observing a scene we used a purposely designed "gaze machine" consisting of a stereo camera and an inertial platform aligned along the optical axis and mounted on a helmet worn by the subject. A further camera was placed on the helmet edge, pointing to the right eye and working as eye-tracker. In this way we were able to store data related to the subjects' head and eye movements, as well as those related to their vantage point.

Three subjects (with normal or correct to normal vision) were asked to observe an environment and to search a not specified number of targets, placed on different planes. The subjects were sitting and could rotate only their head. They were told to fixate briefly every detected target and then to search on for the others. In the process frames from the stereo camera, including depth information, and rotation angles from the inertial platform were recorded conjunctly.

Afterwards a mosaic image was assembled from the RGB frames, in order to project the gaze itinerary upon it.

We will refer in the sequel to an image as a pair of frames $\langle \mathbb{F}_C, \mathbb{F}_D \rangle$, where $\mathbb{F}_C$ is the colour image $m \times n \times 3$, taking values in the range of the three channels *red*, *green* and *blue*, and $\mathbb{F}_D$ is the distance image $m \times n \times 3$, taking values in the range $(-\infty,\infty)$, i.e. to a pixel $[x, y]^T$ will correspond the pixel projection in real world coordinates $[X_w, Y_w, Z_w]$ (the first two coordinates, however, are of no use as they are relative to the point of view of the current frame).

An attention sequence is thus a sequence of pairs $T_{CD} = \langle \mathbb{F}_C, \mathbb{F}_D \rangle_t$ indexed with time.

We supposed that at each instant the stared object was to be found in the center of the current frame. The center coordinates were then translated of a value $k = 15$ pixels that would allow for the distance between the eyes and the camera baseline. Moreover, the coordinates were possibly corrected according to the direction and angle extracted from the eye-tracker. Eventually the attained point shall be referred as a *gaze gate*.

The head angles, i.e. yaw (pan) angle ϕ and pitch (tilt) angle ϑ, along with Z_w were used to estimate $[X_t, Y_t, Z_t]$, the real world coordinates of every gaze gate, by means of the rotation matrices.

Fixations gates were distinguished from gaze shifts on the base of a threshold applied to gaze velocity, obtained similarly as in Koch (2004):

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{pmatrix} = [X_t, Y_t, Z_t] \begin{pmatrix} 0 & -\omega_x & \omega_y \\ \omega_x & 0 & -\omega_z \\ -\omega_y & \omega_z & 0 \end{pmatrix} + \begin{bmatrix} U \\ V \\ W \end{bmatrix}$$

Here $(\omega_x, \omega_y, \omega_z)$ and (U, V, W) denote the rotational and translational velocities, respectively, at time t.

As said above, to appreciate the distribution of gaze shifts over the whole scene the gaze gates of the sequence T_{CD} were projected on a panoramic scene, obtained by mosaicing several frames.

A suitable transformation was defined to pass from real world coordinates to pixel coordinates on the mosaic (taking account of focus and mosaic distortion):

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{bmatrix} \begin{pmatrix} X/Z \\ Y/Z \\ 1 \end{pmatrix} + \alpha \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

Here $\alpha > 0$ is a parameter.

A gaze search model

On 2D images translational shifts along the depth axis are hardly detectable, compared with shift along the x

or y axis. It is therefore crucial to consider the gaze shifting along every dimension to understand why a specific area of the scene attracts the subjects' gaze driven by a visual task.

We could note that, however, the gaze behaviour could not be explained only by a distance criterion. As it could be expected, a great amount of gaze gates were collected in areas with the highest frequency. We assumed, moreover, that further criteria would influence an underlying optimization process, such as the ease at memorizing certain zones or the effect of cluttered regions on recurrent paths. Scene analysis is influenced by classical Gestalt criteria and visual attention is determined both by object and spatial factors (Mozer and Vecera 2005). Space-based attention is responsible for grouping of contiguous locations, whereas an object-based attention allows grouping based on features likely to belong to the same objects. Therefore, we segmented the mosaic according to features like colour, gradient, spatial context and mean location depth, as they help to convey both memorization and spatial representation. For example, fixations were sparse on uniform regions as opposed to cluttered ones. In fact, dense regions require several fixations and shifts between close locations, as targets might be hidden and it might be more difficult to identify them.

A *k-mean* segmentation was produced from colour, gradient and context normalised data. These latter come from a partition of the scene in spatially unified zone, labelled by smoothness order.

The depth data were separately segmented partitioning the scene in sectors and assigning them the corresponding mean depth value.

These segmentations were used to classify areas likely to attract fixations. Hence, we introduced a utility function u: $L \mapsto \mathbb{R}$, L is the utility space according to segmentation criteria. This function assigns high utility values to spots in the scene likely to produce a fixation.

Therefore the expected utility of a vector of gaze gates $X = (x_1, x_2, ..., x_n)$ can be written as:

$$E_u(X) = \sum_{X \in M_D} f_{\vartheta,s}(x) u_{\vartheta,s}(x)$$

where M_D is the depth map, $f_{\vartheta,s}(x)$ is the density of gaze gate x according to parametric zones ϑ and s, u is the utility function. Considering k distance areas (depending on the volume of the scene) we defined $p_\vartheta(x)$ according to the depth map as a Boltzmann distribution, thus considering further gaze gates less likely than closer ones.

Analogously, we define a probability of x to belong to a zone s, $p_s(x)$, according to the area and depth of s. Finally the probability to enter a gate x is:

$$f_{\vartheta,s}(x) \propto \alpha p_\vartheta(x) + (1 - \alpha) p_s(x)$$

Here α is a mixing parameter, consistently defined with the scene depth.

However, while a wide wall could have a high probability due to its surface it might be glimpsed at just once if no target occurs on it. A utility function is thus defined for each criterion. With respect to depth we define $u(z) = k - k\,(z/\beta)$, with β a threshold. With respect to context memorization we take a fixed value γ, and finally with respect to frequency we introduce a value δ proportional to the gradient of the area. Considering that z is given by the depth map, we obtain:

$$u_{\vartheta,s}(x) = u(z) + \gamma + \delta$$

We consider a configuration of attention to be the smallest number of gaze gates maximizing the expected utility. We computed it on the mosaics of the experiments noting a strong correlation between attention and space classification yielded by segmentation based on distance, memorization and frequency criteria.

References

Couyoumdjian A, Nocera FD, Ferlazzo F (2003) Functional representation of 3d space in endogenous attention shifts. Q J Exp Psychol 56A(1):155–183

Frintrop S, Nüchter A, Surmann H (2004) Visual attention for object recognition in spatial 3d data. In: Proceedings of the 2nd international workshop on attention and performance in computational vision (WAPCV), vol. 3368, Prague, Czech Republic, pp 168–182

Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3):194–203

Koch C (2004) Models of motion perception: introduction and the Reichardt correlation model. Notes of the course "vision: from computational theory to neuronal mechanisms", CNS/Bi/EE 186, Caltech

Maringelli F, McCarthy J, Steed A, Slater M, Umiltà C (2001) Shifting visuo-spatial attention in a virtual three-dimensional space. Cogn Brain Res 10:317–322

Mozer M, Vecera S (2005) Object- and space-based attention. In: Itti L, Rees G, Tsotsos J (eds) The encyclopedia of the neurobiology of attention. Elsevier, Amsterdam, pp 130–134

Niebur E, Itti L, Koch C (2001) Controlling the focus of visual selective attention. In: Hemmen LV, Domany E, Cowan J (eds) Models of neural networks IV. Springer, Berlin Heidelberg New York

Ouerhani N, Hügli H (2000) Computing visual attention from scene depth. In: 15th International Conference on Pattern Recognition, vol 1, Barcelona, Spain, pp 375–378
Previc F (1998) The neuropsychology of 3-d space. Psychol Bull 124:123–164
Rensink R, O'Regan J, Clark J (2000) On the failure to detect changes in scenes across brief interruptions. Vis Cogn 7:127–145
Simons D, Levin D (1997) Change blindness. Trends Cogn Sci 1:261–267
Theeuwes J, Atchley P, Kramer F (1998) Attentional control within 3-D space. J Exp Psychol Human Percept Perform 24(5):1476–1475
Treisman A, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12:97–136