EURASIP Journal on Applied Signal Processing 2005:14, 2391–2396

Attentional Mechanisms for Interactive Image Exploration

Joseph Machrouh

Situated Perception Group, Human-Machine Communication Department, LIMSI-CNRS, BP 133, 91403 Orsay, France
Email: joseph.machrouh@limsi.fr

France Telecom Research & Development, 2 Pierre Marzin Avenue, BP 50702, 22307 Lannion Cedex, France
Email: joseph.machrouh@rd.francetelecom.com

Philippe Tarroux

Situated Perception Group, Human-Machine Communication Department, LIMSI-CNRS, BP 133, 91403 Orsay, France

École Normale Supérieure, 45 rue d'Ulm, 75230 Paris Cedex 05, France
Email: philippe.tarroux@limsi.fr

Received 31 December 2003; Revised 15 December 2004

A lot of work has been devoted to content-based image retrieval from large image databases. The traditional approaches are based on the analysis of the whole image content both in terms of low-level and semantic characteristics. We investigate in this paper an approach based on attentional mechanisms and active vision. We describe a visual architecture that combines bottom-up and top-down approaches for identifying regions of interest according to a given goal. We show that a coarse description of the searched target combined with a bottom-up saliency map provides an efficient way to find specified targets on images. The proposed system is a first step towards the development of software agents able to search for image content in image databases.

Keywords and phrases: exploratory vision, bottom-up exploration, top-down exploration, attention, situated vision.

1. INTRODUCTION

Image analysis is confronted with the development of large image databases and new techniques have to be designed for image and content retrieving in this context. The agent paradigm has proved its efficiency for searching in unstructured databases. An agent exhibits interaction abilities with its environment and an autonomous behavior driven by its perceptions of the environment and its expectancies. This viewpoint emphasizes the role of interaction in visual processing and is related to the active vision paradigm mainly used in robotics [1, 2]. We propose here to use a similar paradigm of active vision for implementing content retrieval mechanisms in fixed image or video sequences. To drive the active vision system, we need a mechanism for identifying salient regions in the visual scene. Most of the systems proposed for the computation of saliency maps are based on bottom-up approaches [3, 4]. We use here a bottom-up mechanism to identify a first set of salient regions and a top-down mechanism for target recognition. Salient regions can be defined as high-energy contrast regions. On the other hand, regions of interest are characterized by their high relevance according to a given goal. Preattentional mechanisms are based on saliencies while attentional top-down processes are goal-directed. We thus propose an approach that combines both mechanisms in the following way.

We distinguish two nested regions in an image: the whole visual field, a low-resolution area that can be shifted by attention from position to position, and a small central foveal region that can be analyzed at full resolution. A first set of points is computed at low resolution from the whole visual field and used to give the focus to each potentially interesting region one at a time. We study here how information on the target can bias this exploratory step and improve its efficiency. We also compare different approaches for identifying or rejecting the target when it is foveated.

2. MODEL

2.1. Definition of a saliency space

The first step in our work consisted in defining a projection space in which we can compute the saliencies present in

the visual field. Although saliencies can be computed from various methods (e.g., local image contrast), we assumed here that saliencies are mainly based on preferred orientations and spatial frequencies. Consequently, we used an approach based on Gabor wavelets. The image convolution by a bidimensional Gabor wavelet can be described by the equation

$$r(\mathbf{x}, \mathbf{\Omega}_{k,\theta}) = e^{(-1/2)\mathbf{x}^{\tau}\Sigma^{-1}\mathbf{x}} e^{-i\Omega_{k,\theta}\mathbf{x}} * I(\mathbf{x}), \tag{1}$$

where $I(\mathbf{x})$ is the initial image, $r(\mathbf{x}, \mathbf{\Omega}_{k,\theta})$ the filtered image, and $e^{(-1/2)\mathbf{x}^{\tau}\Sigma^{-1}\mathbf{x}} e^{-i\Omega_{k,\theta}\mathbf{x}}$ is the Gabor convolution kernel. $\mathbf{\Omega}_{k,\theta}$ is a row vector defining the preferred orientations of the filter such that $\mathbf{\Omega}_{k,\theta} = \mathbf{\Omega}_k \mathbf{R}_\theta$ where $\mathbf{R}_\theta$ is the rotation matrix defining the orientation of the filter and $\mathbf{\Omega}_k = (\omega_k \quad 0)$ the central frequency of the filter.

In the present work $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$ and $k \in \{1/12, 1/6, 1/3 \text{ cyc/pixel}\}$.

Thus, starting from the hypothesis that only low frequencies are used to orient the exploratory bottom-up mechanism, we computed the saliency map as explained below.

From a statistically significant set of natural images analyzed through a Gabor wavelet bank, we extracted small image patches at random. Each patch had the same size as the foveal region. From each patch, we computed as many signature vectors $\mathbf{v}_k = \{\bar{r}_{k,\theta}\}_{\theta \in \{0,\pi/4,\pi/2,3\pi/4\}}$ as the number of desired frequency bands according to the following equation:

$$\bar{r}^2_{k,\theta} = \frac{1}{N} \sum_{\mathbf{x}} r(\mathbf{x}, \mathbf{\Omega}_{k,\theta}) \times r^*(\mathbf{x}, \mathbf{\Omega}_{k,\theta}), \tag{2}$$

where N is the number of pixels in the image patch and $r^*(\mathbf{x}, \mathbf{\Omega}_{k,\theta})$ and $r(\mathbf{x}, \mathbf{\Omega}_{k,\theta})$ are complex conjugates.

The multiresolution technique used to compute the $\mathbf{v}_k$ vector is similar to the one proposed by [5]. A principal component analysis (PCA) was then applied to each of these vectors for each spatial frequency channel according to $\mathbf{z} = \mathbf{U}^{\tau}\mathbf{v}$ where $\mathbf{U}$ is an orthogonal projection matrix such that $\langle \mathbf{z}\mathbf{z}^{\tau} \rangle$ is diagonal.

We thus obtained four projection axes in each frequency band, the components of which are linear combinations of the initial orientations. The obtained projection space is significant of the second-order statistical regularities observed in the used subset of natural images. However, experiments performed with various subsets did not show significant differences.

2.2. Preattentional and attentional controls

The saliencies of the scene at each position in the visual field can then be obtained as the projection of the $\mathbf{v}_k$ vectors on the corresponding axis of the PCA (Figure 1). We have shown elsewhere that the salient points computed by this method differ according to the considered axis [6]. Here only the first eigenvector at low resolution was used.

The obtained salient points are used to control the exploration of the scene. In the present study, two methods were used: the bottom-up control uses only information extracted from the visual scene in a Preattentional way, while the top-down control implements an attentional mechanism driven by a previously memorized information concerning the target.

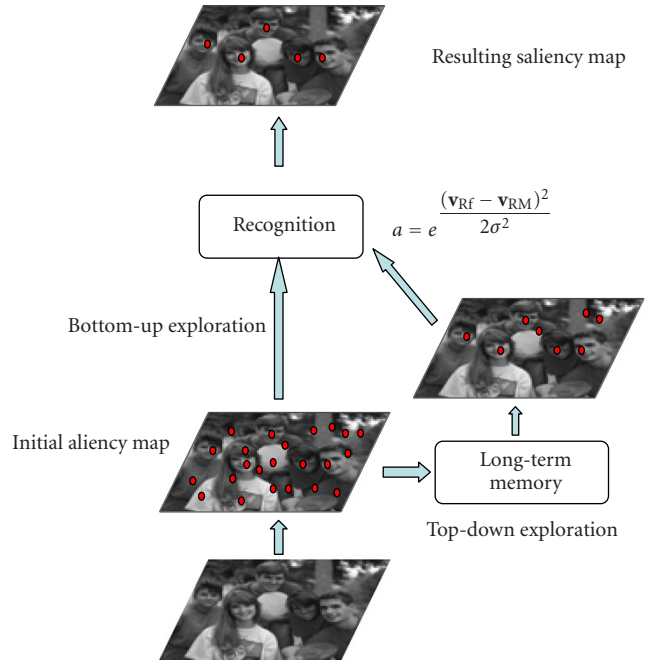


Figure 1: An overall presentation of the attentional model. A bottom-up saliency map is biased with the information on the desired target lying in long-term memory.

We tested this architecture on a task where the system's behavior is to find targets similar to the one pointed out by the user.

In bottom-up control mode, when the user points to a region, the system finds the nearest salient point in its present visual field, focuses on it, and then computes the low-resolution bottom-up salient points in its new visual field. It then focuses on the most salient of these points and computes a recognition score of the target. Two kinds of scores have been tested: (i) one from the average of the Gabor norms, (ii) the other being simply the concatenation of the Gabor norm image vectors covering the foveal area of the system. In this study, these vectors are of dimension 12 (3 spatial frequencies, 4 orientations).

In top-down control mode, the system performs a low-resolution comparison between the salient points in its whole visual field and a low-resolution signature of the target. It thus retains only salient points superior to a given threshold. This mechanism leads to a modulation of the natural saliency of the considered point according to the low-resolution characteristics of the searched target. Two kinds of score computations were tested: (i) a comparison of the energy vectors computed from the low-resolution part of the multiresolution analysis, respectively, from the salient point $\mathbf{x}_s$ and the target representation $\mathbf{x}_t$ (TDE) (ii) a direct comparison of the low-frequency images of the salient region and of the target

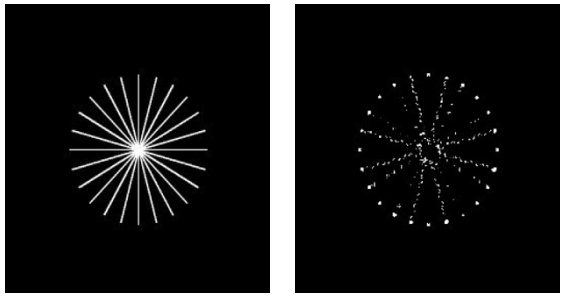


Figure 2: Bottom-up detection of interest points. The figure illustrates the end-stopping (termination detector) properties of the approach.

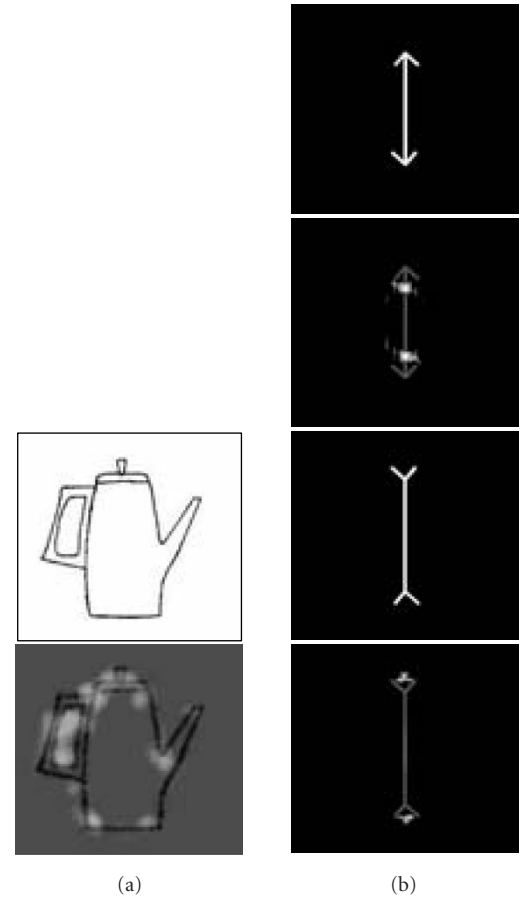
(TDV). The similarity score is thus computed using a radial basis function $a = e^{-\|\mathbf{x}_s - \mathbf{x}_t\|^2/2\sigma^2}$.

2.3. Discussion of the model properties

Some points concerning this approach deserve to be discussed before the description of the obtained results. Major results have been obtained during the last decade concerning the first steps of visual processing in natural systems [7, 8, 9]. These papers show that the first filtering steps consist in the elaboration of an optimal code based on the maximization of a statistical independence criterion. It leads to similar filters such as those obtained using independent component analysis (ICA) [10, 11, 12]. They have been shown to be very similar to Gabor filters [13]. This is why we use this approach in our model.

However, it is interesting to analyze the kind of salient features obtained from the computations described above. Experiments with several different images demonstrated that the features emphasized by such projections mainly consist in termination and curvature points. For instance, some of the features extracted from a test image according to the first PCA axis are rotation-invariant curvature points (Figure 2).

Due to their properties of end-stopping detectors, it is interesting to observe that the saliant positions computed from the image in Figure 3 can be invoked as an explanation for the Müller-Lyer illusion.



(a) (b)

Figure 3: Bottom-up detection of interest points. (a) Detection of interest points is made on the basis of curvature and termination characteristics. (b) The energy peak from these detectors is located inside the direct arrowheads and outside the reversed arrowheads, as expected in the Müller-Lyer illusion where the direct arrowheads appear shorter than the reverse arrowheads.

3. RESULTS

Although the system can be used in various object search tasks, we only present here the results obtained in a face retrieval task.

The user points a face in a scene and the task of the system is to find similar patterns across the image. On this task, we tested the three methods presented above (bottom-up, top-down energy (TDE), and top-down vector (TDV), see Figure 4).

In bottom-up mode, the system is driven by the natural saliencies computed from the scene. These saliencies are sorted according to their decreasing intensities in such a way that the system begins its exploration with the highest intensity saliency. The similarity score obtained in this case ranges from 0.1–1.0. Ten percent of the points have a similarity score in the range 0.9-1.0, while 17% are in the range 0.8-0.9. The majority of the points have a score in the range 0.6–0.9.

In the top-down mode, the system is guided through high-level information. In TDE mode, the similarity scores range from 0.3–1.0. Fourteen percent of the points lie between 0.9 and 1.0 while 22% range from 0.8-0.9. Most of the visited points have a similarity score between 0.7 and 1.0.

In TDV mode, there is a decrease in the variability of the similarity score. Sixty five percent of the points have a similarity score in the range 0.9–1.0 and 10% between 0.8 and 0.9. The most visited points lie between 0.9 and 1.0. The use of top-down information leads to a significant reduction in the number of visited points (234 for the bottom-up exploration, 107 for TDE, and 31 in TDV for the example in Figure 4).

When this experiment is repeated with various images (up to 20 images), faces always yield similarity scores greater than 0.8. We retained this value as a decision threshold separating faces and nonfaces locations. We were thus able to compute an error rate for the different experiments from a

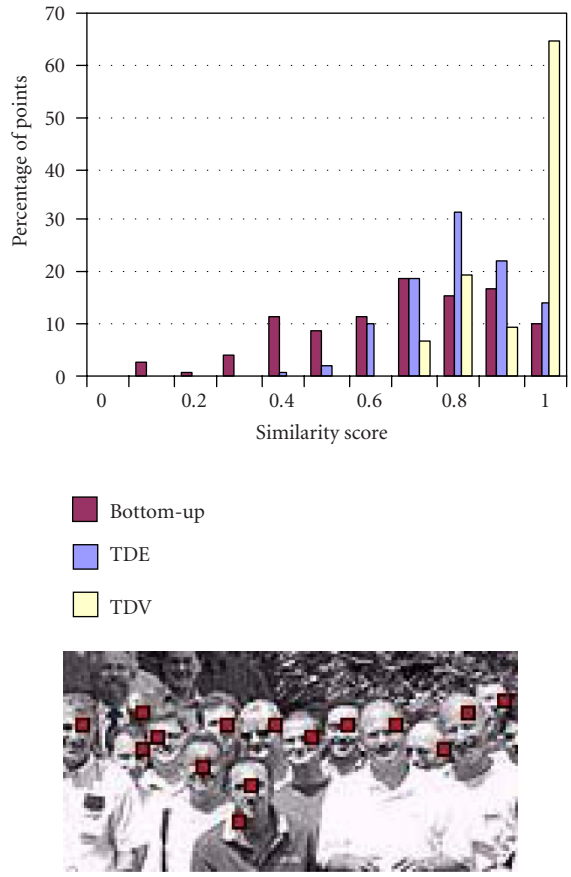


FIGURE 4: Percent of visited points according to the similarity score. The figure shows that a large portion of visited points have a low similarity score in bottom-up exploration, while in TDE and in TDV the visited points exhibit greater similarity scores. The image shows the result obtained with the face recognition task in TDV mode.

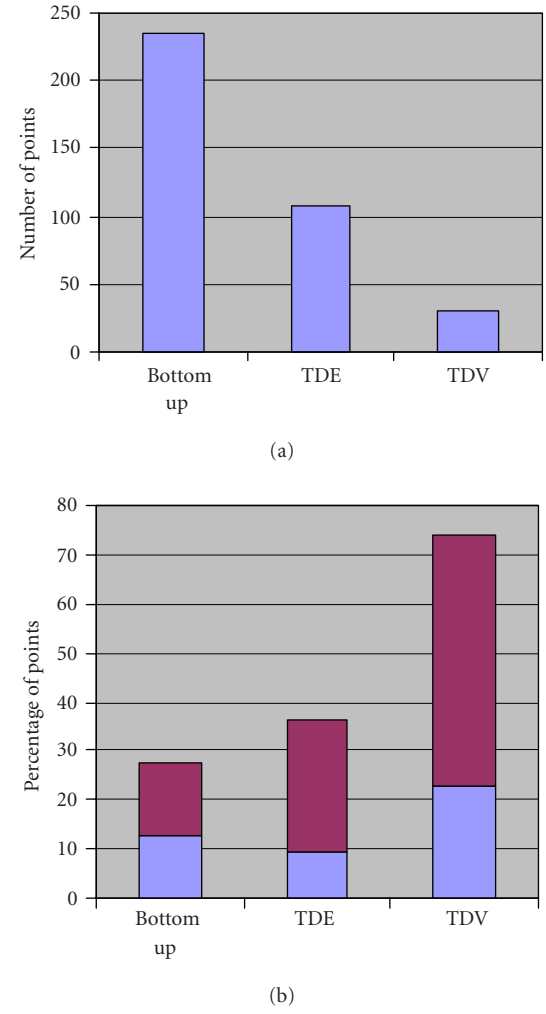


FIGURE 5: (a) Evolution of the number of points explored by the system in the three investigated modes. (b) Evolution of the ratio between faces and nonfaces in the visited points (upper values) and evolution of the recognition error rate (lower values).

comparison between the answer of the system (a similarity score greater than 0.8 being now considered as a positive answer) and the ground truth of the target.

It results from these investigations that in the bottom-up mode only 27% of the visited points are faces while this percentage increases to 36% in the TDE mode and reaches 74% in the TDV mode (Figure 5). On the other hand, in the bottom-up mode the error rate is 47%. It decreases to 26% and 30% in TDE and TDV, respectively. The TDV method gives rise to the best results.

One mandatory specification of this kind of system is its robustness according to the variations of illumination. We tested the behavior of the system in the case of the search for identical targets in a series of video images. This property is indeed especially important in the case where we want to follow the same object through a video sequence. We have used the TDV mode to search for a zone pointed out by the user in a midilluminated scene (image mean intensity 151.9 expressed in grey level) through a set of homologous images the illumination of which ranges from 69.24–185.69. Figure 6a shows the variation of the similarity score according to the illumination for homologous points (i.e., points corresponding to the same target, in order to detect false negatives).

Figure 6b shows the same result for heterologuous points (i.e., points corresponding to different targets, in order to detect false positives). The mean score remains approximately constant in function of illumination. Its variance increases with illumination but the discrimination ability of the system (measured by the threshold between the two curves) is preserved.

4. DISCUSSION AND CONCLUSION

The system presented in this paper is based on two principles: (i) the selection of salient points used to guide exploratory saccades, (ii) the combination of bottom-up and top-down information to bias the saliencies in favor of the searched target. This last modulation reduces the computational load of the system. The identification of the salient points is indeed

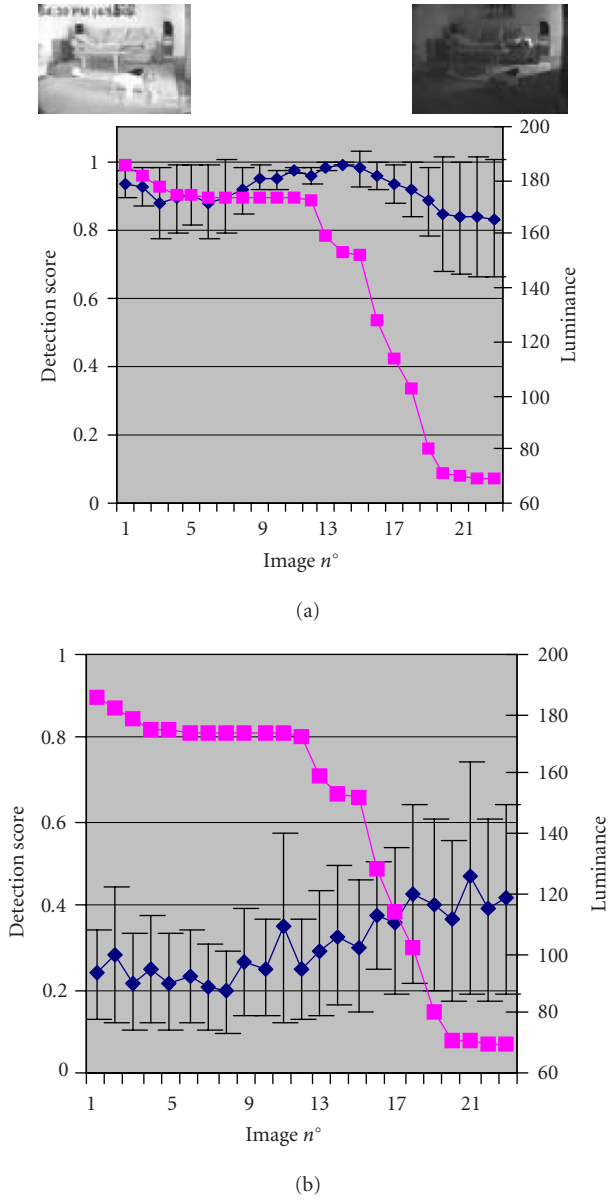


Figure 6: Robustness to the variations of illumination. A video sequence with a continuous variation in luminance has been used to follow the detection of homologous interest points from image to image. The figure shows the mean detection score for (a) targets and (b) nontargets superimposed with the luminance curve (expressed in grey levels).

not based on a saliency map computed on the whole scene [3, 14, 15] but limited to the visual field and computed at low resolution.

The list of the potentially interesting coordinate points can thus be viewed as a sparse representation of the scene consisting of a system of references to the external location where the complete information lies. Such a view was first introduced by O'Regan who proposes to see the world as an external memory [16]. It implements the first principles of the sensori-motor theory of perception proposed by this author [17]. This mechanism is also related to the notion of deictic pointers proposed by Ballard [18]. Note that only stable landmarks can be used for this purpose and that new questions could arise in the case of video applications.

The proposed architecture allows to perform any search and exploration task. It is indeed independent of the type and size of image and the searched target.

Our final goal is to build an exploratory vision architecture able to work in real-time. The reduction of the computational load is critical to achieve this goal. This constraint explains the limited number of preferred directions used in the computation of saliencies and the relative simplicity of the coding method.

The multiresolution technique used here, which performs the complex processing steps on previously selected regions, also provides a mechanism to overcome the real-time constraints. Though the retained information does not allow a complete reconstruction of the initial scene, it is sufficient to ensure a sufficiently fast exploration mechanism. The advantages of this approach, which distinguishes low-resolution and large-field processing from high-resolution focused computations, is twofold. It indeed reduces the need for complex computation during the exploration process and, perhaps more importantly, clearly separates the exploration and exploitation steps that constitute the behavior of the system. As suggested by psychophysics experiments [19], we make the hypothesis that the identification processes happening in peripheral and central vision are quite different. In peripheral vision, we do not need to cope with invariance, since the available representation is simplified, partial, and sparse. It is only made of a set of pointers useful for driving action. From these regions, it seems to be impossible to get a complex representation of objects [19]. On the contrary, the central part of the visual field provides the information for building complex objects representations. One of the major contributions of the proposed approach is that the system does not need a complete representation of the object to select locations to focus at. The recognition process can thus take place in two steps: (i) identification of potentially interesting locations according to the searched target, (ii) recognition of the target after foveation. When the search process is biased by low-resolution information related to the target, the number of potentially interesting points dramatically decreases which improves the efficiency of the search process. We can thus parallel this mechanism with the one at work in natural vision system in which the search for a given target could be driven by a simplified description of the target, the recognition process being made easier by the fact that it operates only on focused regions.

One can argue that the proposed method is neither rotation- nor scale-invariant. However, it is inherently invariant in translation; since the targets will eventually be centered, the translational invariance problem disappears.

Another interesting fallout of considering perception as a dynamical mechanism is that the system endowed with those perceptual abilities can be viewed as a kind of autonomous agent. The interactive process in which the agent is involved can thus be improved using learning techniques

popular within the agent's or robotics communities. Among these methods, the use of reinforcement learning is presently under investigation in our laboratory.

REFERENCES

[1] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," in *Proc. 1st International Conference on Computer Vision*, pp. 35–54, London, UK, 1987.

[2] R. Bajcsy, "Active perception," *Proc. IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[3] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[4] J.-M. Bost, R. Milanese, and T. Pun, "Temporal precedence in asynchronous visual indexing," in *Proc. 5th International Conference on Computer Analysis of Images and Patterns (CAIP '93)*, pp. 468–475, Springer Verlag, Budapest, Hungary, September 1993.

[5] A. Guérin-Dugué and P. M. Palagi, "Texture segmentation using pyramidal Gabor functions and self-organising feature maps," *Neural Processing Letters*, vol. 1, no. 1, pp. 25–29, 1994.

[6] Y. Machrouh, J. S. Lienard, and P. Tarroux, "Multiscale feature extraction from visual environment in an active vision system," in *Proc. International Workshop on Visual Form 4, Capri, Italy*, Springer-Verlag, Berlin, Germany, May 2001.

[7] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America* {*A*}, vol. 4, no. 12, pp. 2379–2394, 1987.

[8] D. J. Field, "What is the goal of sensory coding?" *Neural Computation*, vol. 6, no. 4, pp. 559–601, 1994.

[9] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[10] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[11] J. Hérault, A. Oliva, and A. Guérin-Dugué, "Scene categorization by curvilinear component analysis of low frequency spectra," in *Proc. 5th European Symposium on Artificial Neural Networks (ESANN '97)*, pp. 91–96, Bruges, Belgium, April 1997.

[12] A. Guérin-Dugué and H. Le Borgne, "Analyse de scènes naturelles par composantes indépendantes," in *Ecole de printemps. de la séparation de sources à l'analyse en composantes indépendantes, Méthode, algorithmes et applications*, Villard-de-Lans, France, June 2001.

[13] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[14] R. Milanese, "Detecting salient regions in an image: from biological evidence to computer implementation," Ph.D. dissertation, University of Geneva, Geneva, Switzerland, 1993.

[15] L. Itti, C. Gold, and C. Koch, "Visual attention and target detection in cluttered natural scenes," *Optical Engineering*, vol. 40, no. 9, pp. 1784–1793, 2001.

[16] J. K. O'Regan, "Solving the 'Real' mysteries of visual perception: the world as an outside memory," *Canadian Journal of Psychology*, vol. 46, no. 3, pp. 461–488, 1992.

[17] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioral and Brain Sciences*, vol. 24, no. 5, pp. 939–973, 2001.

[18] D. H Ballard, M. M. Hayhoe, P. K. Pook, and R. P. N. Rao, "Deictic codes for the embodiment of cognition," *Behavioral and Brain Sciences*, vol. 20, no. 4, pp. 723–724, 1997.

[19] M. Boucart, M. Fabre-Thorpe, S. Thorpe, C. Arndt, and J. C. Hache, "Covert object recognition at large visual eccentricity," *Journal of Vision*, vol. 1, no. 3, pp. 471–472, 2001.

Joseph Machrouh received his M.S. degree in mathematics and computer science from Paris 1 University. From 1998 to 2002, he was with the LIMSI-CNRS and received his Ph.D. degree in 2002 from Paris XI University. He is presently in a postdoctoral position with France Telecom Research & Development. His research interests include attentive vision, face detection, and face and gesture tracking.

Philippe Tarroux graduated from the École Normale Supérieure, Paris. He received his Ph.D. degree from Paris University in 1984 in the field of natural systems modelling. At the École Normale Supérieure he was in charge of the Bioinformatics Group until 1990 and was one of the cofounders of the AnimatLab. He joined the LIMSI-CNRS in 1998 where he developed research on situated perception. He is presently in charge of the Man-Machine Communication Department, LIMSI-CNRS, and his research interests are focused on attentional artificial vision and perception-action relationships in autonomous robots and artificial entities.