Perceptual agents: A situated framework for image analysis

Joseph Machrouh[1], Philippe Tarroux[1, 2]
Situated Perception Group
[1] LIMSI-CNRS, BP 133 F-91403 Orsay France
[2] ENS, 45 rue d'Ulm, F-75230 Paris Cedex 05 France
{joseph.machrouh, philippe.tarroux}@limsi.fr

Abstract

In this paper, we propose an algorithm for computer vision that reconciliates the two main approaches used up to now in the field. A large part of the image processing and understandings methods rely on straightforward algorithmic approaches, but robotic applications and artificial intelligence have inspired methods based on exploratory behavior and active vision. In this paper, we show that it is possible to design a system able to perform complex tasks on an image or a video sequence by the means of exploratory techniques usually developed for computer vision without incurring a too high computing cost.

1. Introduction

Classical image analysis and computer vision techniques have been designed to extract the content of an image or a visual scene using a reconstructionnist approach. Computer vision is thus grounded on the mentalist conception of cognition. Another approach has emphasized the role of interaction in the study of visual processes [1]. This viewpoint has lead to the active vision paradigm of robotics [2] [3]. We propose here to use the same paradigm of active vision for the analysis of fixed images or video sequences. We thus introduce the concept of perceptual agent, a software agent designed to actively search information in image and video sequences databases. These exploratory mechanisms give rise to interesting algorithms and the use of an agent architecture allows more adaptive modes of interaction. To drive an active vision system, we need a mechanism to identify salient regions in the visual scene. Most of the algorithms proposed for the computation of saliency maps are bottom-up [4] [5]. We propose here to identify a first set of interest points using such a bottom-up mechanism and to use a top-down one for target recognition. The set of points computed at low resolution over the whole visual field is used to give the focus to each potentially interesting region one at a time. The top-down information is then used within the focused region to identify or reject putative targets. We thus propose solutions to the following questions:

- How to compute the points of interest in a visual scene?
- How to combine the information coming from the scene with the memory content of the system and its internal expectancies?

We show that when the search process is biased by low-resolution information related to the target, the number of potentially interesting points dramatically decreases and the efficiency of the search process improves. This approach leads to an efficient algorithm of target selection due to the low computational cost of the exploratory phase. We can parallel this mechanism with the one at work in natural vision systems in which the search for a given target could be driven by a simplified description of the target, the recognition process being made easier because it operates only on focused regions.

2. Model

The interest points are defined as high-energy regions. They are computed through a bottom-up filtering process using a bank of Gabor filters described by:

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

where $I(\mathbf{x})$ is the initial image, $r(\mathbf{x},\Omega_{k,\theta})$ the filtered image and $e^{-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}}$ the Gabor kernel used to convolve the image.

$\Omega_{k,\theta}$ is a vector defining the preferred orientations of the filter such that $\Omega_{k,\theta} = \Omega_k\mathbf{R}_\theta$ where $\mathbf{R}_\theta$ is a rotation matrix and $\Omega_k = (\omega_k \quad 0)$.

In the present work $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$ and $k \in \{1/32, 1/16, 1/8, 1/4\}$

Only low frequencies are used to orient the exploratory bottom-up mechanism, while the complete frequency range is retained for the definition of the goal to be retrieved in top-down exploration.

The second step in the computation of the saliencies is the extraction of higher-level characteristics from the output of the Gabor filters. As stated above, the system is based on a recognition mechanism running once focused on a region of interest. Therefore, the potential target is always focused when recognition occurs. Thus, we have to compute saliencies as if the targets were always centred. The suitable method to obtain an optimal code is to use Independent Component Analysis (ICA) [6] [7] [8]. Several authors have shown the efficiency of Principal Component Analysis under the hypothesis that the images are centred [9] [10] [11].

We extract random small image patches from a statistically significant set of natural images. Each patch has the same size as the foveal region. From each patch, we compute as many signature vectors $\mathbf{v}_k = \{\bar{r}_{k,\theta}\}$ as the number of frequency bands according to the following equation

$$\bar{r}_{k,\theta} = \overline{\left\| r(\mathbf{x},\Omega_{k,\theta}) \right\|} = \frac{1}{m \times n} \sum_{\mathbf{x}} r(\mathbf{x},\Omega_{k,\theta}) r^*(\mathbf{x},\Omega_{k,\theta})$$

where m and n are the row and the column numbers of the patches.

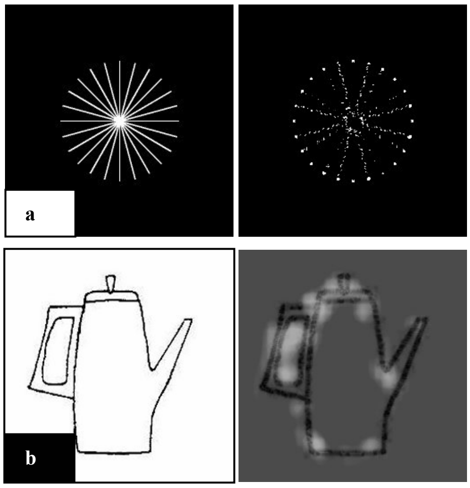

Figure 1. Bottom-up detection of interest points. The figure shows that the detection of interest points is made on the basis of curvature and termination characteristics.

A PCA was 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.

The multi-resolution technique used to compute the $\mathbf{v}_k$ vectors is based on a Gaussian pyramid and is similar to the one proposed by [12]. We thus obtain four vectors for each frequency band. The resulting projection space is significant of the statistical regularities observed in the subset of natural images used here. Experiments performed with various subsets do not show significant differences. The saliencies are computed for each position in the visual field as the projection of the $\mathbf{v}_k$ vectors on the corresponding axis of the PCA. We have shown in [13] that the salient points computed by this method differ according to the considered axis. In this study only the first eigen-vector at low-resolution are used.

Further studies are necessary to determine more precisely the nature of the features emphasized by such projections. Experiments with several images demonstrate that these features mainly consist in termination and curvature points. Some of the features extracted from a test image according to the first PCA axis are rotation-invariant curvature points (Fig 1).

These salient points are used to control the exploration. We use two methods:

(i) The bottom-up control uses only information extracted from the visual scene in a pre-attentional way.

(ii) The top-down control implements an attentional mechanism driven by a previously memorized information on the target.

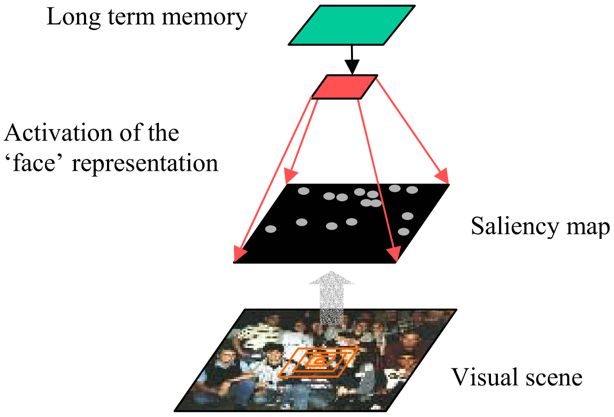

Figure 2. The system modulates the natural saliency of the considered point according the low-resolution characteristics of the searched target.

We tested this architecture to find the targets similar to the one pointed by the user. When the user points to a region, the system finds the nearest salient point, focus on it and computes the low-resolution bottom-up salient points in its visual field. It then focus on the most salient of these points and computes a new vector $\mathbf{v}$ representing a complete description of the target at this point in terms of orientations and spatial frequencies. This vector is used to compute the recognition score of the target. Two descriptions have been tested, one from the average of the Gabor norms, 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 mode, the system performs a low-resolution comparison to retain only salient points superior to a given threshold. It modulates the natural saliency of the considered point according to the low-resolution characteristics of the searched target (figure 2). Two kinds of comparison algorithms were tested:

(i) a comparison of the energy vectors computed from the low-resolution part of the multi-resolution analysis respectively from the salient point and the target representation

(ii) a direct comparison of the low-frequency images of the salient region and of the target.

These comparisons are computed using a radial basis function $s = e^{-\frac{\|\mathbf{v}-\mathbf{w}\|^2}{2\sigma^2}}$. The output of this function gives a similarity code.

3. Results

3.1. Exploration

In this experiment, we tested these techniques on a face identification 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)).

In the 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 intense saliency. The similarity score obtained in this case range from 0.1 to 1.0. 10% of the points have a similarity score in the range 0.9-1.0, while 17% are in the range 0.8-0.9. Most of the points have a score in the range 0.6-0.9.

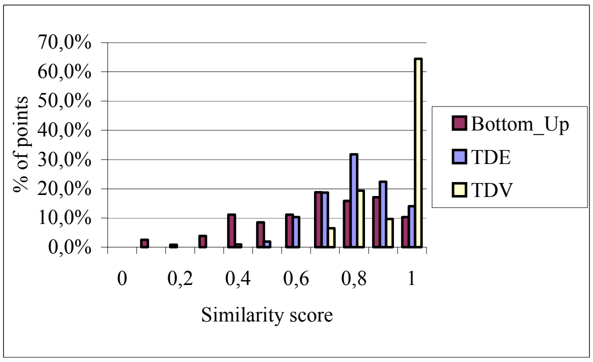

Figure 3. Percent of visited points according to the similarity score. The figure shows that a large portion of visited points have a low similarity score in bottom-up exploration while in TDE and in TDV, the visited points exhibit greater similarity scores. The image shows the result obtained with the face recognition task in TDV mode.

In top-down mode, the system is guided through high-level information. In TDE mode, the similarity scores range in 0.3-1.0. 14% of the points lie between 1.0 and 0.9 while 22% range in 0.9-0.8. Most of the visited points have a similarity score between 0.7 and 1.0 (figure 3).

In TDV mode, there is a drop in the variability of the similarity score. 65% 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 a 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 image Figure 4).

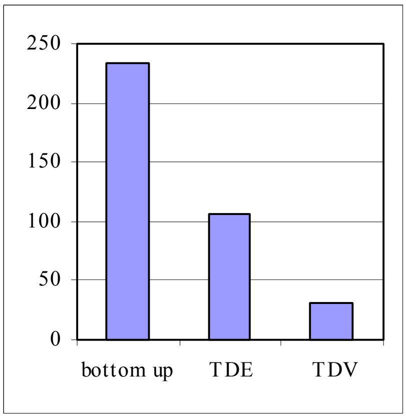

Figure 4. The evolution of the number of points explored by the system in the three investigated modes.

When this experiment is repeated with various images (up to 20 images), faces always had similarity score greater than 0.8. We thus decided to adopt this value as a decision threshold separating faces and non-faces locations. Consequently, we can compute an error rate for the different experiments from a comparison between the answer of the system (a similarity score greater than 0.8 being now considered as a positive answer) and the real nature of the target.

It results from these investigations that only 27% of the visited points are faces in the bottom-up mode while this percentage drops out to 36% in the TDE mode and reaches 74% in the TDV mode. 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 (Figure 5). 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 behaviour 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 we want to follow the same object through a video sequence. We have used the TDV mode to search for a zone pointed by the user in a mid-illuminated scene

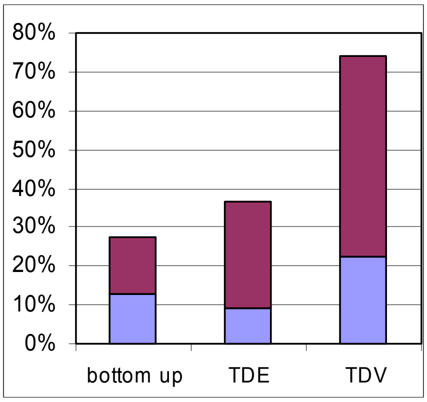

Figure 5. Evolution of the ratio between faces and non-faces in the visited points (upper values) and evolution of the recognition error rate (lower values).

(image mean intensity 151.9 expressed in grey level) through a set of homologous images the illumination of which ranges form 69.24 to 185.69

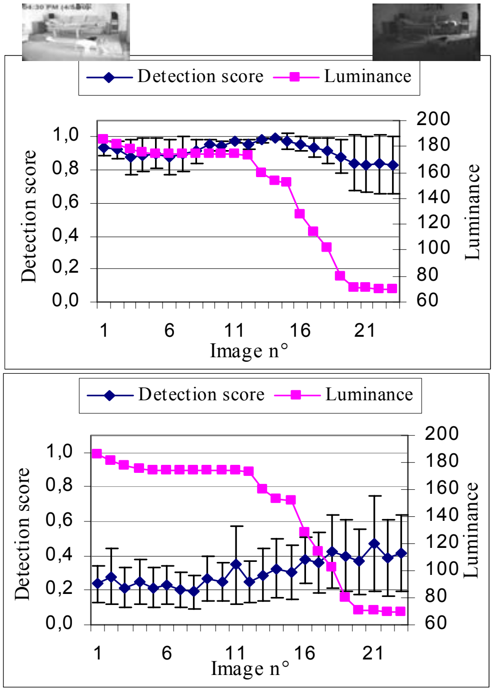

Figure 6. Robustness to the variations of illumination. A video sequence with a continuous variation in luminance has been used to follow the detection of homologous interest points from image to image. The figure shows the mean detection score for target (left) and non target (right) superimposed with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

Figure 6 left 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 6 right shows the same result for heterologous 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 here is a first step toward the implementation of software agents performing more complex task on images. In its present form it 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 favour of the searched target. This last modulation reduces the computational load of the system. The identification of the salient points is indeed not based on a saliency map computed on the whole scene [14] [4] [15] but limited to the visual field and computed at low-resolution. The proposed architecture allows to perform any search and exploration task. It is indeed independent of the type and size of image and of the searched target.

The system retains only the potentially interesting points. In this sense, it works on a sparse representation of the scene consisting in an index of the interesting locations. The full information corresponding to these locations is never coded into the memory of the system. It is retrieved from the internal reference, the world itself been used as an external memory [16]. Note that it is only adapted to the use of stable landmarks. It could raise new questions in the case of video and robotics applications. However, it implements the first principles of the sensori-motor theory of perception proposed by O'Regan and Noë [17]. This mechanism also relates to the notion of deictic pointers proposed by Ballard and col. [18]. The two step search procedure based on low-frequencies first and on a full representation in a second time, implements a kind of hypothesis verification mechanism, the identification of a target being viewed as a reasoning procedure.

Our goal in this study was to build an exploratory vision architecture able to work in real-time. This constraint explains the limited number of preferred directions used and the relative simplicity of the coding method. Though the retained information does not allow a complete reconstruction of the initial scene, it is sufficient to allow a correct exploration mechanism.

The reduction of the computational load is critical to achieve this goal. The multi-resolution technique used here, which performs the complex processing steps on previously selected regions, provides the mechanism to overcome these constraints. 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 to complex computation for the exploration process and perhaps more importantly clearly separates the exploration and exploitation steps that constitute the behaviour of the system.

The proposed method is not scale-invariant. However, the coding method is inherently invariant in translation. We showed that the identified saliencies are rotation-invariant. Thus, they can be used for matching 3D views of complex objects characterised by a set of local features [19]. However, additional work is required to use this approach for object recognition and to compare it to more global approaches [20].

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 representation available is simplified, partial and sparse. It is only made of a set of pointers useful for driving action and, in biological system, the most evolutionary primitive part of the visual system. From these regions, it seems to be impossible to get a complex representation of objects. On the contrary, the central part of the visual field provides the information for building complex object representations. However, since the targets are centred, the translational invariance problem disappears.

The approach presented here points out that image analysis can be viewed as an active process and that, far from increasing the complexity of the problem, this dynamical perspective helps find out solutions based on a form of reasoning actively using image information.

Another interesting fallout to consider systems endowed with those abilities is that they can be viewed as autonomous agents. 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.

5. References

[1] P.E. Agre and D. Chapman, "Pengi: an implementation of a theory of activity", *AAAI*, 1987, pp. 268-272

[2] Y. Aloimonos, I. Weiss and A. Bandyopadhyay, "Active Vision", *1st International Conference on Computer Vision*, 1987, pp. 35-54

[3] R. Bajcsy, "Active Perception", *IEEE*, 76 (8), 1988, pp. 996-1005

[4] L. Itti and C. Koch, "A Saliency-based search mechanism for overt and covert shifts of visual attention", *Vision Research*, 40, 2000, pp. 1489-1506

[5] J.-M. Bost, R. Milanese and T. Pun, "Temporal Precedence in Asynchronous Visual Indexing", *4th International Conference on Computer Analysis of Images and Patterns*, Springer Verlag, 719, 1993

[6] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications", *Neural Networks*, 13, 2000, pp. 411-430

[7] J. Hérault, A. Oliva and A. Guérin-Dugué, "Scene categorisation by curvilinear component analysis of low frequency spectra", *ESANN'97*, Bruges, 1997, pp. 91-96

[8] A. Guérin-Dugué and H. Le Borgne, "Analyse de scènes naturelles par composantes indépendante", *Ecole de printemps. De la séparation de sources à l'analyse en composantes indépendantes, Méthode, algorithmes et applications*, Villard-de-Lans, 2001, pp. 279-289

[9] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces", *Journal of the Optical Society of America*, 4, 1987, pp. 519-524

[10] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the charcterization of humans faces", *IEEE trans. Pattern Analysis and Machine Intelligence*, 12 (1), 1990, pp. 103-108

[11] M. Turk and A. Pentland, "Eigenfaces for recognition", *Journal of Cognitive Neuroscience*, 3 (1), 1991, pp. 71-86

[12] A. Guérin-Dugué and P.M. Palagi, "Texture segmentation using pyramidal Gabor functions and self-organising feature maps." *Neural Processing Letters*, 1 (1), 1994, pp. 25-29

[13] Y. Machrouh, J.S. Lienard and P. Tarroux, "Multiscale feature extraction from visual environment in an active vision system", *International Workshop on Visual Form 4*, Springer Verlag, Berlin, Capri, It, 2001, pp. 388-397

[14] R. Milanese, "Detecting Salient Regions in an Image from Biological Evidence to Computer Implementation", *Department of Computer Science*, University of Geneva, Switzerland, 1993, pp. 176

[15] L. Itti, C. Gold and C. Koch, "Visual attention and target detection in cluttered natural scenes", *Optical Engineering*, 40 (9), 2001, pp. 1784-1793

[16] J.K. O'Regan, "Solving the "Real" Mysteries of Visual Perception: The World as an Outside Memory", *Canadian Journal of Psychology*, 46 (3), 1992, pp. 461-488

[17] J.K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness", *Behavioral and Brain Sciences*, 24 (5), 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*, 20 (4), 1997, pp. 723

[19] D.G. Lowe, "Local feature view clustering for 3D recognition", *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2001, pp. 682-688

[20] A. Leonardis and H. Bischof, "Robust recognition using eigenimages", *Computer Vision and Image Understanding*, 78 (1), 2000, pp. 99-118