Available online at www.sciencedirect.com
ScienceDirect
Image and Vision Computing 25 (2007) 597–613
www.elsevier.com/locate/imavis

Dynamic visual attention model in image sequences

María T. López [a,b], Miguel A. Fernández [a,b], Antonio Fernández-Caballero [a,b,*], José Mira [c], Ana E. Delgado [c]

[a] Departamento de Sistemas Informáticos, Escuela Politécnica Superior de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain
[b] Instituto de Investigación en Informática de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain
[c] Departamento de Inteligencia Artificial, E.T.S.I. Informática, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain

Received 10 September 2004; received in revised form 2 May 2006; accepted 16 May 2006

Abstract

A new computational architecture of dynamic visual attention is introduced in this paper. Our approach defines a model for the generation of an active attention focus on a dynamic scene captured from a still or moving camera. The aim is to obtain the objects that keep the observer's attention in accordance with a set of predefined features, including color, motion and shape. The solution proposed to the selective visual attention problem consists in decomposing the input images of an indefinite sequence of images into its moving objects, by defining which of these elements are of the user's interest, and by keeping attention on those elements through time. Thus, the three tasks involved in the attention model are introduced. The Feature-Extraction task obtains those features (color, motion and shape features) necessary to perform object segmentation. The Attention-Capture task applies the criteria established by the user (values provided through parameters) to the extracted features and obtains the different parts of the objects of potential interest. Lastly, the Attention-Reinforcement task maintains attention on certain elements (or objects) of the image sequence that are of real interest.

Keywords: Dynamic visual attention; Motion; Segmentation; Feature extraction; Feature integration

1. Introduction to selective attention

Findings in psychology and brain imaging have increasingly suggested that it is better to view visual attention not as a unitary faculty of the mind but as a complex organ system sub-served by multiple interacting neuronal networks in the brain [33]. At least three such attentional networks, for alerting, orienting, and executive control, have been identified. The images are usually built from the entries of parallel ways that process distinct features: motion, solidity, shape, color, location [9]. One of the most influential theories about the relation between attention and vision is the Feature-Integration Theory [37]. Treisman hypothesized that simple features were represented in parallel across the field, but that their conjunctions could only be recognized after attention had been focused on particular locations. Recognition occurs when the more salient features of the distinct feature maps are integrated.

The first neurally plausible architecture of selective visual attention was proposed by Koch and Ullman [25], and is closely related to the Feature-Integration Theory. In [23], a visual attention system inspired by the behavior and the neural architecture of the early primate visual system is presented. The MORSEL (Multiple Object Recognition and attentional SELection) model [31] links visual attention to object recognition in order to provide an explicit account of the interrelations between these two processes. In [21], a neural network (connectionist) model called the Selective Attention for Identification Model (SAIM) is introduced. SAIM can model a wide range of experimental evidence on normal attention and attentional disorders [20]. The model of Guided-Search (GS) by Wolfe [41] uses

* Corresponding author. Tel.: +34 967 599200; fax: +34 967 599224.
E-mail addresses: mlopez@info-ab.uclm.es (M.T. López), miki@info-ab.uclm.es (M.A. Fernández), caballer@info-ab.uclm.es (A. Fernández-Caballero), jmira@dia.uned.es (J. Mira), adelgado@dia.uned.es (A.E. Delgado).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2006.05.004

the idea of saliency map to realize the search of objects in scenes. It is able to find one item in a visual world filled with other distracting items. In [8] a system of interconnected modules consisting of populations of neurons for modeling the underlying mechanisms involved in selective visual attention is proposed. In [34], the SCAN (Signal Channelling Attentional Network) architecture is presented. The building block of SCAN is a gating lattice, a sparsely connected neural network. SCAN introduces a biological solution to the problem of translation-invariant pattern processing. In [38], a model that is able to obtain objects separated of the background in static images is presented. Thus, the previous models have in common that they provide explanations for a wide range of experiments on normal and abnormal visual perception and attention, and they introduce neurally inspired architectures applied to static images.

On the other hand, some visual attention models have shown their interest in including motion analysis. A recent model of attention for dynamic vision has been introduced by Backer and Mertsching [1]. In this model there are two selection phases. Previous to the first selection a saliency map is obtained as the result of integrating the different features extracted. In particular, the extracted features are symmetry, eccentricity, color contrast, and depth. The first selection stage selects a small number of items according to their saliency integrated over space and time. These items correspond to areas of maximum saliency and are obtained by means of dynamic neural fields. The second selection phase has top-down influences and depends on the system's aim (''behind'', ''higher'' or ''larger'', and so on). In [18,28] NAVIS (Neural Active VIsion System), object recognition is performed in a multistage way starting from the hypothesis of the presence and localization of an object. Then, it identifies the object from its parts. Features extracted at the bottom-up process are axis orientation, areas orientation, color and motion. The static features are combined jointly with the top-down information of the presence of an object to perform the recognition process.

An excellent survey of approaches to computational attention is given in [35]. The interest paid to attention has grown a lot recently and is being used in real-world applications. Some implemented systems based on selective attention have so far covered up several of the following categories: recognition [4,19,32,36,40], teleconferencing [22], tracking of multiple objects [5,39], and mobile robot navigation [2,3,42,29,6].

2. In search of relevant adaptive selective attention parameters for segmentation

Our approach defines a model for the generation of an *Active (Dynamic) Attention Focus* on a dynamic scene. The aim is to obtain the objects that keep the observer's attention in accordance with a set of predefined features, including color, motion and shape. In other words, the used features are related to the motion and shape of the elements present in the grey-level images dynamic scene. Thus, our proposal follows an attentional-scene-segmentation-integrating approach [27], where shape and motion are integrated. The model may be used to observe real environments indefinitely in time (there is no limit in the number of input images) with the purpose of tracking a wide variety of objects. In relation to the most common motion suppositions described in [30], our approach obtains good results, as (i) objects need not stay in the scene, (ii) our method does not impose any restriction on null or constant motion of the camera, (iii) more than just one single object may capture the attention in the scene, and (iv) our proposal deals any kind of motion. Let us insist on the fact that our computational model performs well with static and moving cameras. In relation to environmental suppositions, that is to say, constant illumination, static image background, and uniform background, we can state, without any doubt, that our model is a good one.

Fig. 1 shows the result of applying our model to the generic *Dynamic Visual Selective Attention* task, where attention has been paid on moving elements belonging to the ''car'' class.

The solution proposed to the selective visual attention problem consists in decomposing the input images of an indefinite sequence of images into its moving objects, defining which of these elements are of the observer's – or user's – interest, and keeping attention on those elements through time. In the system proposed it is mandatory that the observer may define the features of the objects on which attention is focused. The commands (or indications) that the observer introduces into the system in order to adjust parameters which define the attention focus are of a top-down modulation. This modulation is included in a static way during the process of feature election, as well as in a dynamic form established as a feedback from the attention focus where parameters which define the interest may be modified to centre the focus on objects that are of real interest.

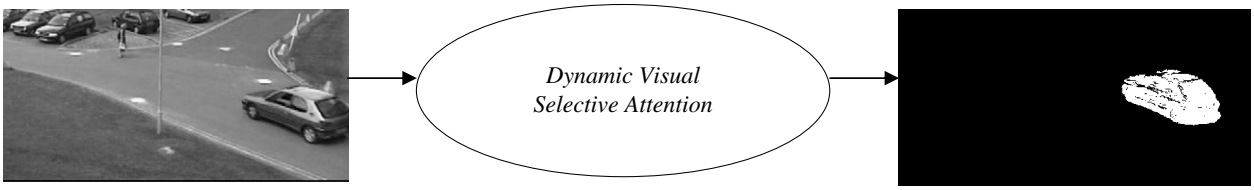

Fig. 1. Input and output of the ''Dynamic Visual Selective Attention'' task.

Our solution defines a model including two kinds of processes: bottom-up processes, where pixel and object features are extracted, and top-down processes, where the observer organizes mechanisms and search parameters to satisfy his expectations with respect to the attention focus.

The selection of the elements of interest in the scene necessarily starts with setting some criteria based on features extracted from the elements (*Feature Extraction*). First, all the necessary mechanisms to provide sensitivity to the system are included in order to succeed in centering the attention. Frame to frame it will be possible to capture attention (*Attention Capture*) on elements made up from image pixels that fulfill the requirements established by the user. On the other hand, stability has been provided to the system. This has been gotten by including mechanisms to reinforce attention (*Attention Reinforcement*), in such a way that the elements that assemble the user's predefined requirements are strengthened up to be configured as the system attention centre. Fig. 2 shows the decomposition into subtasks of the generic *Dynamic Visual Selective Attention* task. In this figure, the three previously introduced subtasks are depicted:

- *Feature Extraction*: Obtains those features (color, motion and shape) of the image able to capture attention.
- *Attention Capture*: Applies the criteria established by the user (values provided to parameters) to the extracted features and obtains the different parts of the objects of potential interest.
- *Attention Reinforcement*: Maintains attention on certain elements, or objects, of the image sequence that are of real interest.

Moreover, the solution to the problem takes into account two additional basic factors:

- First, the nature and characteristics of the input signal has been largely studied. Thus, in this paper the problematic complexity underlying image processing on indefinite video sequences coming from real scenarios is described.

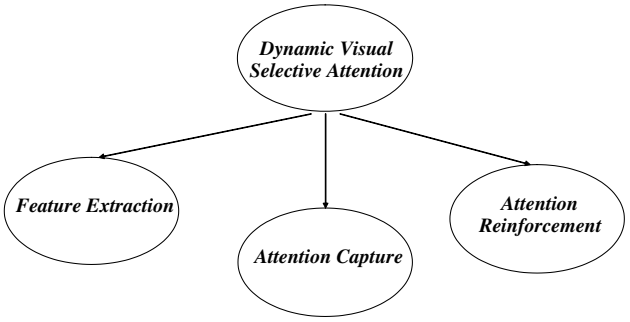

Fig. 2. Subtasks of generic "Dynamic Visual Selective Attention" task.

- Second, there is the facet corresponding to the observer's intentions (or interests). This paper includes all the necessary mechanisms to attend the user's commands that allow directing and keeping attention on scene objects that fulfill a series of predefined features. At each image frame a new attention focus is calculated; it may be the same one as in the previous frame, or it may change to another object or objects depending on the changing observer's desires.

Next each one of these factors is analyzed in detail.

2.1. The input signal

The input information comes from a real scenario where diverse objects are moving through time in a long-range series of images. These objects continually change in their shapes as well as in their three-dimensional spatial position. Typical examples of such scenarios are traffic scenes, visual surveillance scenes, and so on.

The first problem in image processing is related to the conversion of the real scene to information accessible to the computation world. This conversion is performed by means of the recording of the scene on a physical support giving rise to the digitalization. In general, there are two possibilities in relation to image capture: a recording with one single camera, or a recording with more than one camera (stereoscopy in the case of two cameras). In this work, images from a single camera are used. The image is made up of a determined number of pixels distributed as a three-dimensional matrix formed by columns, rows and time.

It is well known that the problems associated with image processing are vast and complex [17,24]. Next some of the fundamental problems are mentioned, as they have played an important role in the evolution of the mechanisms that have led to the final solution proposed in this paper. In first place, and in relation to the problems due to digitalization, let us highlight the problems derived from spatial and temporal sampling. Discreet sampling imposes a series of limitations to motion as well as to the size of the scene objects that the system may process. Second, passing from a three-dimensional scenario to a two-dimensional image, added to the fact of facing moving scenes, also throw their associated range of problems. The most important of them may be the fact that even rigid objects in a three-dimensional scene appear in our image as deformable elements due to the variations in position or orientation through time or to occlusions with other objects. There are also the proper problems associated with motion, such as the well-known aperture problem, the normal changes in illumination, and the correspondence problem.

In short, there is very little invariant input information. Candidates to centre the attention focus have no invariants in shape, in size and generally even in their illumination levels or chromaticity. The matter of which decision features are appropriate for the user is not trivial.

Next some of the features related to the input image that may be used to face the problem of selective visual attention are described. Among this set of features, we will comment which of them have been used in the *Dynamic Visual Selective Attention* system proposed. The features regarding the input signal have been classified into image pixel features and scene element (or object) features.

2.1.1. Features related to image pixels

The only feature of a pixel of coordinates (x,y) of a two-dimensional image obtained from the recording of a scene in time instant t is the quantified value of the electric signal. In the case of monochrome images, there are usually 256 grey levels. In some digital image processing applications the number of levels is even decreased. The most extreme level diminishment that still provides information on the image consists in using only two levels; that is to say, there will be a binary image.

Considering the spatial relations of a pixel of coordinates (x,y) with respect to its neighborhood environment, some information might be obtained by applying several convolution masks. Applying such masks provides discontinuities, isolated pixels, lines, borders, etc. Now, considering temporal relations the application of filters enables to obtain information on illumination variation, motion detection, etc.

2.1.2. Features related to image objects

Features related to objects are those that contain information referred to scene elements. Scene elements are to be understood as a set of connected image pixels that hold a series of common features. On the one hand, there are features associated with the shape of the objects, such as size, width, height, width–height ratio, eccentricity, compactness and similarity to standard shapes (matching). On the other hand, there is the information related to the chromatic features, such as grey level, color or brightness. There are also features associated to the motion of the elements, such as velocity, acceleration, length-speed ratio [10–12], and so on.

2.1.3. Other features

There is another family of parameters that may be used to direct the attention focus, namely those features associated to an extra-contextual behavior. That is to say, we are looking for pixels or objects – see also parts of objects – with different features from the rest, as, for instance, different grey level, different texture, different velocity, different shape or different motion features. The problem of defining the pixels of interest can be faced from different perspectives. In some cases, as, for example, in models based only on the scene, the interest pixels are obtained by distinguishing their features from those of the surrounding pixels [23].

2.1.4. Selected features

Some of the features pointed out in the previous sections might have been used alone or combined to select potentially interesting pixels or objects to the observer. It is convenient to highlight that features associated to pixels are easier to obtain than features associated to elements. In the latter case an additional problem arises, namely the one of segmenting the objects present in the scene [7,13,43,44].

Thus, obtaining features associated with image pixels as well as scene segmentation into different elements are fundamental to the solution adopted in this paper. Another important cue is maintaining the attention focus on a particular object in a sensitive and stable way. Our research team's prior knowledge of the problem has led to the selection the following features from the ones previously described:

- At pixel level the chosen features are grey level, motion detection, velocity and acceleration. This decision is supported by the fact that these features have largely been considered as interesting in lots of applications [10–16] where selection and segmentation are crucial.
- At object level the selection has been much more complicate. In our case, we eventually decided to chose easily computable features and at the same time providing a great capacity of classification. In particular, we have worked with features related to the shape of the objects, such as, for instance, the size of the objects (number of pixels that compose the object), the width of the objects (difference among the highest and lowest row of the object in the image), the height (difference among the highest and lowest column of the object in the image), the width–height ratio (width/height) and the compactness (size/(width * height)).

This set of features is valid for the observer to propose his search intentions and for the system to process information in an efficient way to focus the attention on objects that are really of interest to the user. Once the selected features have been described, we can decompose the *Feature-Extraction* task into the subtasks shown in Fig. 3, namely:

- *Color Feature Extraction*, at image pixel level.
- *Motion Feature Extraction*, also at pixel level of indefinite image sequences.

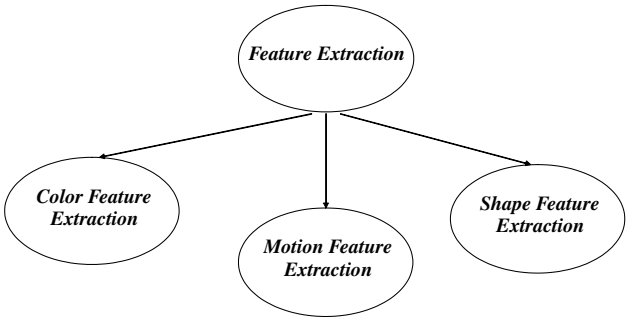

Fig. 3. Subtasks of "Feature-Extraction" task.

- *Shape Feature Extraction*, at image object part (spots or zones of the objects) or total object level.

2.2. The observer

From the observer point of view, the aim of the proposed model is to heed his intentions, which will determine the features of interest at any time. This way one or several objects in the scene that respond to the given parameters are obtained by segmenting scene objects in a continuous way. Once the focus has been centered on particular objects, other information systems may use the results obtained and dedicate to the knowledge extraction tasks, learning, classification, etc.

The observer commands fix the limits of the different feature values. The intentions on the attention focus may be to maintain the focus on the scene object, to abandon the attention or to expand the focus around the object. More precisely, the observer may use one of the following possibilities to indicate his desires: (1) to centre attention on the image object that currently holds the attention focus, (2) to augment the attention focus around the object that configures the attention focus, (3) to abandon attention on the object that currently maintains the attention focus, and (4) to centre attention on image objects that fulfill a combination of the following characteristics: (a) moving image objects, (b) image objects that contain pixels at given velocities, (c) image objects that contain pixels moving at defined acceleration values, (d) image objects with a given size, and (e) image objects with a given shape.

3. Segmentation of moving objects through dynamic visual attention

Once the nature of the problem has been analyzed from the input signal and the observer viewpoints, and attending to their restrictions of capacities, the solution to the problem is now introduced. The proposed solution defines a model with two kinds of processes: bottom-up processes (based on the scene) to extract scene pixel and object features and top-down processes, which enable the observer to manage the search mechanisms and parameters to satisfy his expectations with respect to the attention focus. In Figs. 4 and 5 two schemes of the proposed solution are offered. Particularly, in Fig. 4 there is a description of the subtasks controlled by the observer from the *Commands Generation* subtask. In Fig. 5 you have the general schema of the proposed solution where the subtasks that appeared in Fig. 4 are highlighted:

- *Motion Feature Extraction*,
- *Color Feature Extraction*,
- *Shape Feature Extraction*,

which satisfy the task called *Feature Extraction*, as well as:

- *Attention Capture* and
- *Attention Reinforcement*.

Next all subtasks are described by means of a simple running example shown in Fig. 6. The example consists of a scene where a vehicle (a car) and a pedestrian are

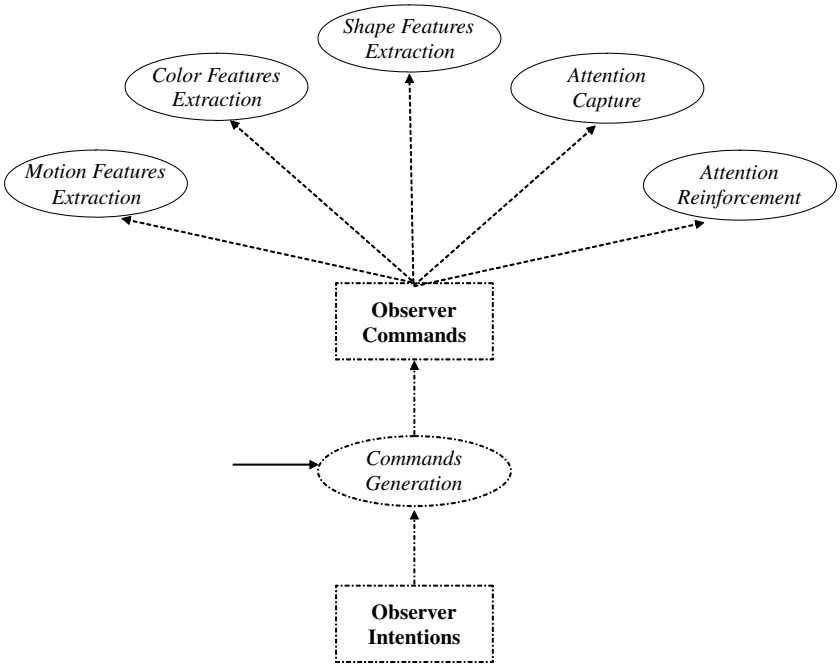

Fig. 4. Subtasks controlled by the observer's intentions.

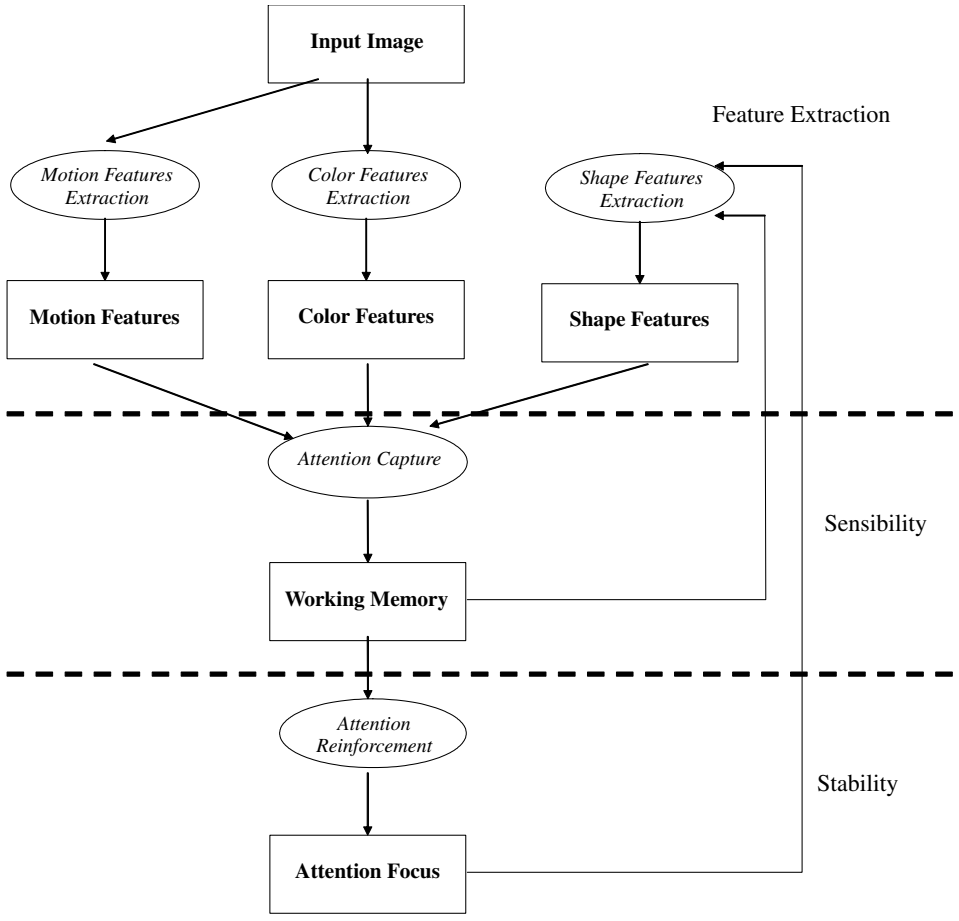

Fig. 5. Schema of the solution proposed.

Fig. 6. Input sequence. (a) First image. (b) Last image.

moving. The sequence is a subset of a test database of the *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance* called "*PETS2001 Datasets, Dataset 1: Moving people and vehicles*", and has been downloaded via ftp://pets.rdg.ac.uk/PETS2001/. The aim of our running example is to maintain attention on those objects that fulfill a series of conditions of size and of dynamics. In the offered example the dynamic conditions are simply the existence of motion with respect to the previous time instant, whereas the size conditions are those of the object "car".

3.1. Color feature extraction

The aim of the *Color Feature Extraction* subtask, as shown in Fig. 7, is to get the chromatic features associated to the image pixels. We work with 256 grey-level input images and transform them to a lower number of levels. Good results are usually obtained with eight levels. These eight-level images are called images segmented into eight grey-level bands (*GLBs*).

Let $GL[x,y,t]$ be the grey level of a pixel (x,y) of the input image at time instant t, $GL_{\max}$ the maximum grey-level value (generally, 255), $GL_{\min}$ the minimum grey-level value (generally, 0), n the number of grey-level bands, and, $GLB[x,y,t]$ the grey-level band of pixel (x,y) at t. Let also S be the overlap (or minimum value of the difference in the grey levels between two consecutive time instants required

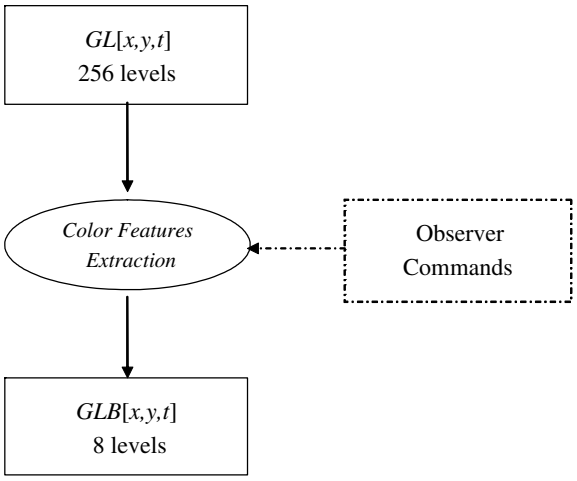

Fig. 7. Subtask "Color Feature Extraction".

to produce a change in the grey-level band of a pixel). Then:

$$GL_{\text{diff}} = GL_{\max} - GL_{\min} + 1 \tag{1}$$

$$GLB[x,y,t] = \begin{cases} GLB[x,y,t-1] & \text{if } \max\left(\frac{(GLB[x,y,t-1]-1)*GL_{\text{diff}}}{n} - S, GL_{\min}\right) \\ & \leqslant GL[x,y,t] < \min\left(\frac{GLB[x,y,t-1]*GL_{\text{diff}}}{n} + S, GL_{\max}\right) \\ \left\lfloor \frac{GL[x,y,t]*n}{GL_{\text{diff}}} \right\rfloor + 1 & \text{otherwise} \end{cases} \tag{2}$$

Eq. (2) checks if grey-level value $GL[x,y,t]$ produces a variation of band in relation to the grey-level band value obtained at $t-1$, that is to say, $GLB[x,y,t-1]$. For this aim, the criteria used is to check if $GL[x,y,t]$ has sufficiently changed in its grey level between time instants t and $t-1$ (use of overlap S). The result is 0 if $GL[x,y,t]$ is in the range established, and 1 in the other case. In Fig. 8 the parts belonging to each one of the eight bands of one of the images in the running example are shown.

3.2. Motion feature extraction

The aim of the *Motion Feature Extraction* subtask is to calculate the dynamic (motion) features of the image pixels, that is to say, in our case, the presence of motion, the velocity and the acceleration. Due to our experience we know some methods to get that information.

Remember that to diminish the effects of noise due to the changes in illumination in motion detection, variation in grey-level bands at each image pixel is performed. Motion presence $Mov[x,y,t]$ is easily obtained as a variation in grey-level band between two consecutive time instants t and $t-1$:

$$Mov[x,y,t] = \begin{cases} 0 & \text{if } GLB[x,y,t] = GLB[x,y,t-1] \\ 1 & \text{if } GLB[x,y,t] \neq GLB[x,y,t-1] \end{cases} \tag{3}$$

Velocity and acceleration are obtained by calculating their respective modules and angles. We start from the memorization along time (accumulation) [10,11] of charge $Ch_{Mov}[x,y,t]$ at each image pixel (x,y). Notice that charge $Ch_{Mov}[x,y,t]$ stores motion information as a quantified value.

$$Ch_{Mov}[x,y,t] = \begin{cases} Ch_{\min} & \\ \quad \text{if } Mov[x,y,t] = 1 & \\ \min(Ch_{Mov}[x,y,t-1] + C_{Mov}, Ch_{\max}) & \\ \quad \text{if } Mov[x,y,t] = 0 & \end{cases} \tag{4}$$

Eq. (4) shows how charge at pixel (x,y) gradually increases through time (frame to frame) in a quantity C_{Mov} (charge constant due to motion) up to a maximum charge or saturation $Ch_{\max}$, while motion is not detected. At the opposite, charge decreases to a minimum of charge $Ch_{\min}$, when motion is detected at pixel (x,y). To calculate the module $|\vec{v}[x,y,t]|$ and the angle $\beta[x,y,t]$ of the velocity, first, velocities in directions x, $v_x[x,y,t]$, and y, $v_y[x,y,t]$, at each pixel are computed.

$$v_x[x,y,t] = \frac{C_{Mov}}{Ch_{Mov}[x,y,t] - Ch_{Mov}[x+1,y,t]} \tag{5.1}$$

$$v_y[x,y,t] = \frac{C_{Mov}}{Ch_{Mov}[x,y,t] - Ch_{Mov}[x,y+1,t]} \tag{5.2}$$

$$\beta[x,y,t] = \arctan\frac{v_y[x,y,t]}{v_x[x,y,t]} \tag{5.3}$$

$$|\vec{v}[x,y,t]| = \sqrt{v_x[x,y,t]^2 + v_y[x,y,t]^2} \tag{5.4}$$

In a similar way, the module $|\vec{a}[x,y,t]|$ and the angle $\alpha[x,y,t]$ of the acceleration at each image pixel (x,y) are computed from accelerations in directions x, $a_x[x,y,t]$, and y, $a_y[x,y,t]$:

$$a_x[x,y,t] = \frac{C_{Mov} \cdot (v_x[x,y,t] - v_x[x+1,y,t])}{Ch_{Mov}[x,y,t] - Ch_{Mov}[x+1,y,t]} \tag{6.1}$$

$$a_y[x,y,t] = \frac{C_{Mov} \cdot (v_y[x,y,t] - v_y[x,y+1,t])}{Ch_{Mov}[x,y,t] - Ch_{Mov}[x,y+1,t]} \tag{6.2}$$

$$\alpha[x,y,t] = \arctan\frac{a_y[x,y,t]}{a_x[x,y,t]} \tag{6.3}$$

$$|\vec{a}[x,y,t]| = \sqrt{a_x[x,y,t]^2 + a_y[x,y,t]^2} \tag{6.4}$$

Fig. 9 shows the result of calculating the presence of motion in our running example. In this concrete case, this is the only dynamic feature that has been indicated by the observer. In the output of this subtask (see Fig. 9), a pixel drawn in white color means that there has been variation in the grey-level band of the pixel in instant t with respect to the previous instant $t-1$.

3.3. Attention capture

The objective of the *Attention-Capture* subtask is to select image zones (or patches) included in objects of interest. It has been decided to construct these patches from image pixels that fulfill the requirements established by

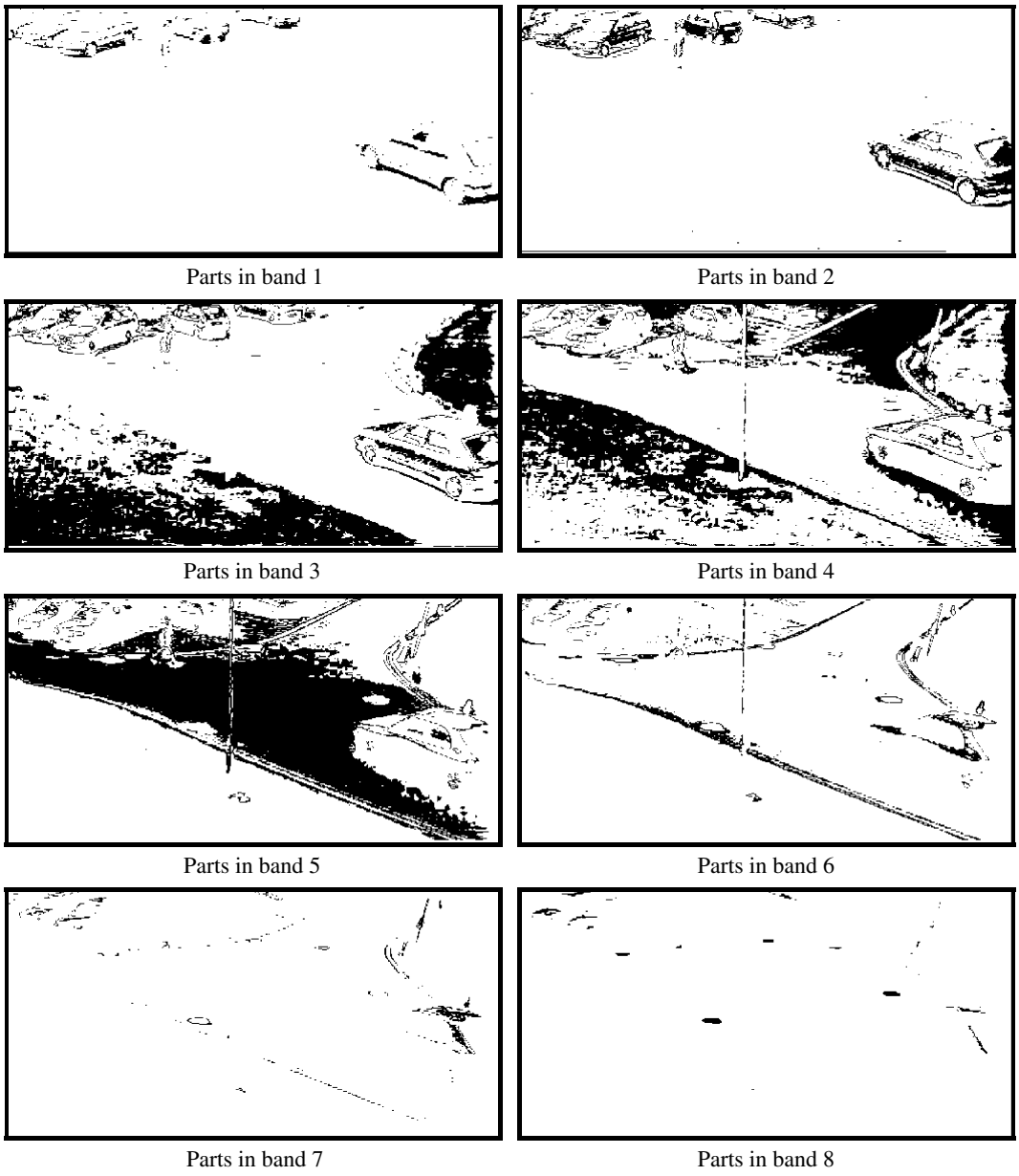

Fig. 8. Image elements' parts in grey-level bands.

the observer's commands. In Fig. 10, a scheme of the *Attention-Capture* subtask is given.

The output of this subtask has been called *Working Memory*. The term *Working Memory* has been chosen due to the similarity with the same concept used in Psychology, where the working memory, also called functional or short-term memory, stores and processes during a brief time the selected information coming from the sensorial paths. In our case, only those elements which appear in the *Working Memory* will potentially convert into the system's attention focus.

Some research lines to solve the problem of defining what are the elements that decompose the scene [38] are based on border extraction and obtain complex objects from more simple ones by looking for families of shapes. Our approach starts by obtaining the object's parts from their grey-level bands. Later on these objects parts (also called zones, patches or spots) will be treated as whole objects.

In previous papers from our research team some algorithms for the segmentation of the image in different objects have been proposed based on the detection of motion, the permanency effect and lateral interaction [13,26]. Thus, based on the satisfactory results of the algorithms commented, we propose, in order to solve the current problem, to incorporate mechanisms of charge and discharge (based on the permanency effect), as well as mechanisms of lateral interaction. These mechanisms are good enough to segment the scene into moving objects and background.

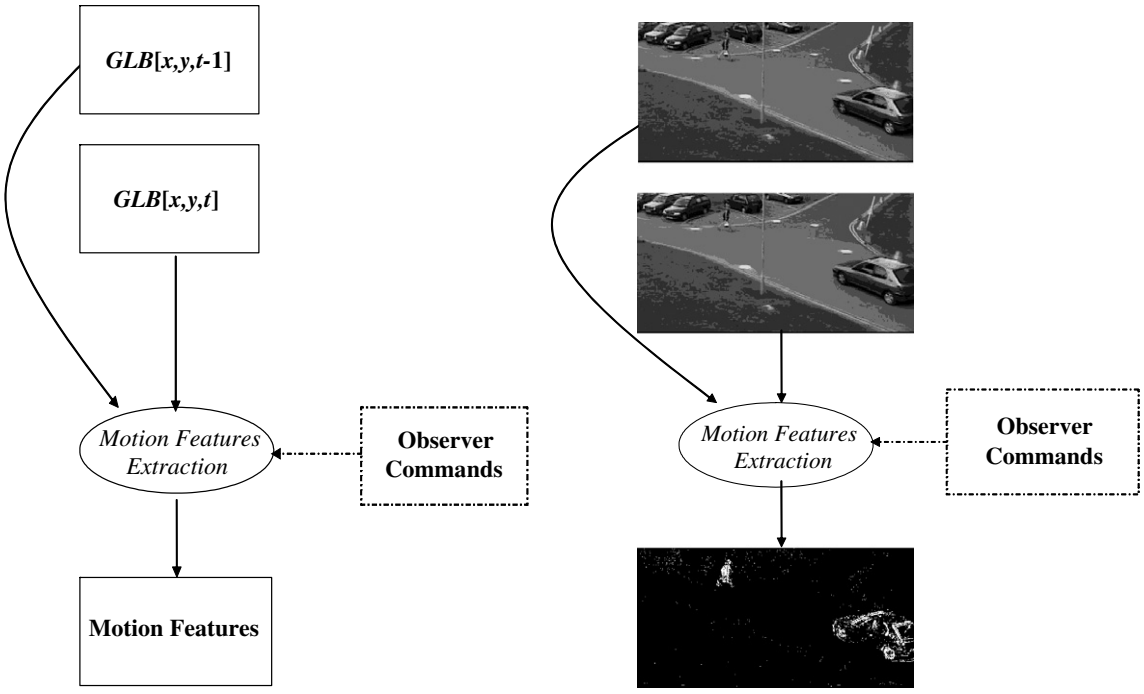

Fig. 9. Subtask "Motion Feature Extraction".

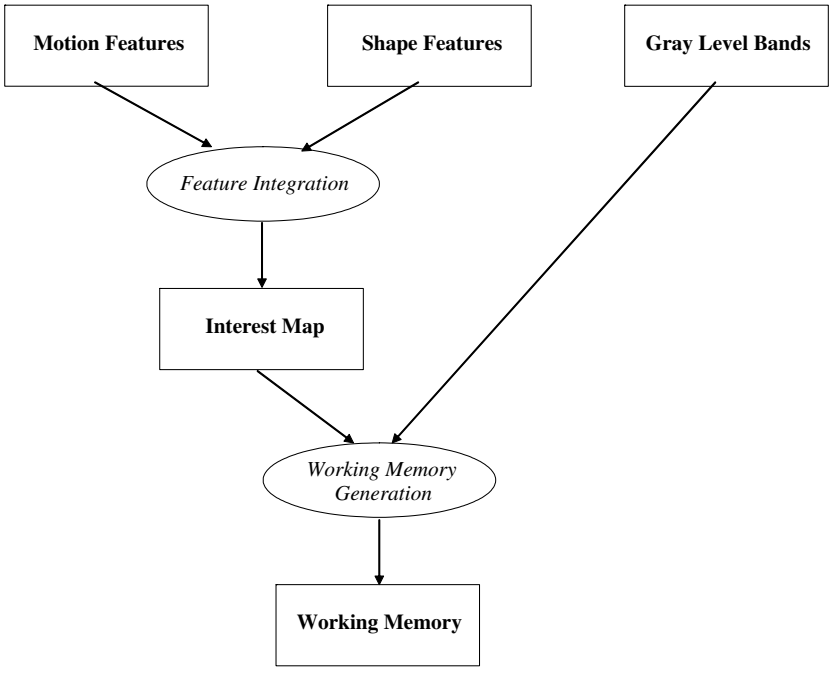

Fig. 10. Subtask "Attention Capture".

In this proposal, the patches present in the *Working Memory* are constructed from the so-called *Interest Map*. The *Interest Map* is obtained, as it will be seen later on, by performing a *Feature Integration* of pixel motion and spot shape features. Spot shape features are those concerned with scene object parts. Thus, first, we will introduce how patches present in the *Working Memory* are obtained from the *Interest Map*. Then, we will explain how *Feature Integration* is performed.

3.3.1. Working memory generation

The aim of this subtask is to construct object spots from image pixels that possess the requirements established by the observer. First, the image is segmented into *Grey-Level Bands* in regions composed of connected pixels whose illumination level belongs to a common interval (grey-level band). Second, only those connected regions that include an "active" pixel in the *Interest Map* are selected. Each one of these regions (or silhouettes) of a uniform grey-level

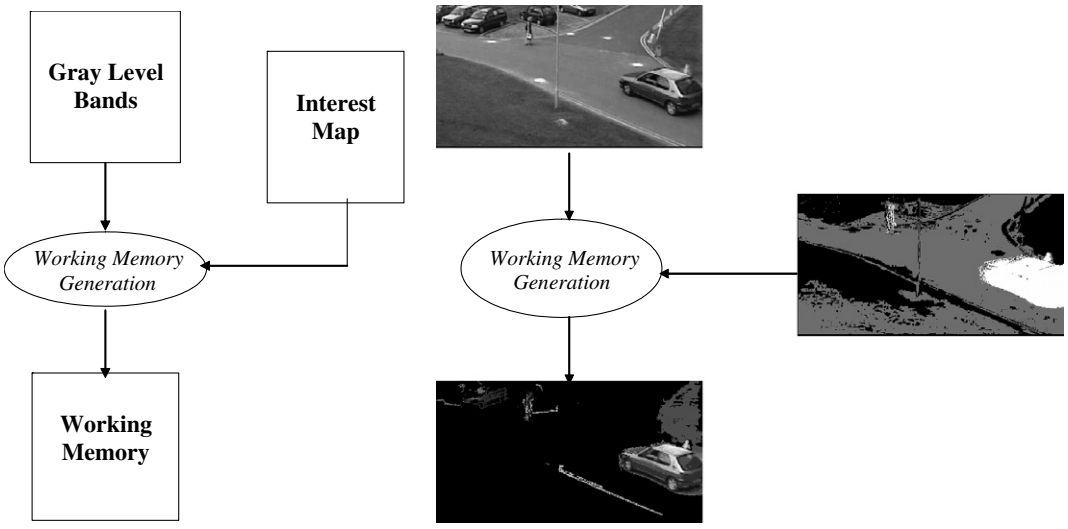

Fig. 11. Subtask "Working Memory Generation".

band is defined as a scene spot belonging to a potentially interesting object.

In order to obtain the patches that contain "active" pixels in the *Interest Map*, the process consists in overlapping, just as done with superimposed transparencies, the image segmented in grey-level bands of the current frame (at t) with the image of the *Interest Map* constructed in the previous frame (at $t-1$).

In Fig. 11, the inputs and the output of the *Working Memory Generation* subtask are shown. The inputs are the image in *Grey-Level Bands*, $GLB[x,y,t]$, and the *Interest Map*, $IM[x,y,t]$. The *Interest Map* contains "active" pixels (in white color), "neutral" pixels (in black color), and "inhibited" pixels (in a different grey-level color). The output of the subtask is the *Working Memory*, $WM[x,y,t]$. The *Working Memory* stores for each pixel belonging to a selected spot a number given to the spot (the label of the spot). Value 0 is for the rest of the pixels, that is to say, to pixels that do not belong to a patch of interest.

As the model works with n grey-level bands, the value at each pixel of the *Working Memory* will be the maximum value of the *Working Memory* calculated at each grey-level band:

$$WM[x,y,t] = \arg\max_i WM_i[x,y,t] \quad \forall i \in [1..n] \tag{7}$$

Next the way in which the *Working Memory* is obtained for each grey-level band is explained. The initial value (patch label) for each pixel (x,y) at grey-level band i is the pixel's position within the image (coordinate x multiplied by the number of image columns + coordinate y) whenever the pixel is in "active" state in the *Interest Map*. A maximum value ($v_{\max}$ = number of columns * number of rows + 1) is assigned if the pixel is labeled as "neutral" and a minimum value ($v_{\min}=0$) if the pixel is "inhibited".

$$v_i[x,y] = \begin{cases} (x * NC + y) + 1 & \text{if } BNG[x,y,t] = i \wedge IM[x,y,t] = v_{\text{active}} \\ v_{\max} & \text{if } BNG[x,y,t] = i \wedge IM[x,y,t] = v_{\text{neutral}} \\ v_{\min} & \text{otherwise} \end{cases} \quad \forall i \in [0..n] \tag{8}$$

This initial value is compared to the neighbors' values that are at the same grey-level band i in an iterative way up to reaching a common value for all the pixels of a same element:

$$v_i[x,y] = \begin{cases} v_{\min} & \text{if } v_i[x,y] = \min(v_i[\alpha,\beta]) = v_{\min} \\ \min(v_i[\alpha,\beta]) & \text{if } v_{\min} < \min(v_i[\alpha,\beta]) < v_i[x,y] \leqslant v_{\max} \\ v_i[x,y] & \text{if } v_{\min} < v_i[x,y] < \min(v_i[\alpha,\beta]) \leqslant v_{\max} \\ v_{\max} & \text{if } v_i[x,y] = \min(v_i[\alpha,\beta]) = v_{\max} \\ & \quad \forall [\alpha,\beta] \in [x \pm 1, y \pm 1] | 0 < v_i[\alpha,\beta] \leqslant v_{\max} \end{cases} \tag{9}$$

Finally, the value established by consensus is assigned to the *Working Memory* at each grey-level band:

$$WM_i[x,y,t] = \begin{cases} 0 \\ \quad \text{if } (v_i[x,y] = v_{\min}) \vee (v_i[x,y] = v_{\max}) \\ v_i[x,y] \\ \quad \text{otherwise} \end{cases} \tag{10}$$

3.3.2. Feature integration

The output of the *Feature Integration* subtask is the *Interest Map* obtained by integrating *Motion Features* and *Shape Features*. The *Interest Map* stores for each image pixel the result of the comparison with three discrepancy classes: "active", "inhibited" and "neutral". This classification is

performed following the observer's commands or intentions. The states of "active" or "inhibited" are reserved for those pixels where motion presence has been detected at current time t (information available in *Motion Features*), or for pixels belonging to an object – or object spot – of interest at time instant $t-1$ (information found in *Shape Features*). Now, "neutral" pixels are the rest of the image pixels. "Active" pixels are those that fulfill the requirements imposed by the user, whilst "inhibited" pixels do not fulfill the requirements.

$$IM[x,y,t] = \begin{cases} v_{\text{active}} & \text{if "discrepancy class 1"} \\ v_{\text{inactive}} & \text{if "discrepancy class 2"} \\ v_{\text{neutral}} & \text{if "discrepancy class 3"} \end{cases} \quad (11)$$

In the running example which we are using to explain our solution to the problem, let us assume that the criteria used by the observer are to "activate all moving vehicles" and to "inhibit all objects that are not cars". The output should be as shown in Fig. 12, white color for "active" pixels, grey color for "inhibited" pixels and black color for the rest of pixels ("neutral").

3.4. Shape feature extraction

The *Shape Feature Extraction* subtask extracts different shape features of the elements stored in the *Working Memory*, $WM[x,y,t]$, (the size $s\_WM[v_{label}]$, the width $w\_WM[v_{label}]$ and the height $h\_WM[v_{label}]$). Let us remember that the labels in the *Working Memory* have been obtained by grey-level bands. Thus, as we have already explained, a moving object is formed by a set of spots with different labels. We call this the *Spot Shape Feature Extraction*.

$$s\_WM[v_{label}] = \text{count}(WM[x,y,t]) | WM[x,y,t] = v_{label} \quad (12.1)$$

$$w\_WM[v_{label}] = \max(y) - \min(y) | WM[x,y,t] = v_{label} \quad (12.2)$$

$$h\_WM[v_{label}] = \max(x) - \min(x) | WM[x,y,t] = v_{label} \quad (12.3)$$

In a similar way, the features of the objects stored in the *Attention Focus*, $AF[x,y,t]$, are obtained (the size $s\_AF[v_{label}]$, the width $w\_AF[v_{label}]$, the height $h\_AF[v_{label}]$, the width–height ratio $hw\_AF[v_{label}]$ and the compactness $c\_AF[v_{label}]$). These are now complete objects united by a common identifying label. So, let us talk about an *Object Shape Feature Extraction*.

Fig. 12. Output of subtask "Feature Integration".

$$s\_AF[v_{label}] = \text{count}(AF[x,y,t]) | AF[x,y,t] = v_{label} \quad (13.1)$$

$$w\_AF[v_{label}] = \max(y) - \min(y) | AF[x,y,t] = v_{label} \quad (13.2)$$

$$h\_AF[v_{label}] = \max(x) - \min(x) | AF[x,y,t] = v_{label} \quad (13.3)$$

$$hw\_AF[v_{label}] = \frac{h\_AF[v_{label}]}{w\_AF[v_{label}]} \quad (13.4)$$

$$c\_AF[v_{label}] = \frac{s\_AF[v_{label}]}{h\_AF[v_{label}] * w\_AF[v_{label}]} \quad (13.5)$$

3.5. Attention reinforcement

The mechanisms used to generate the *Working Memory* endow the system with sensitivity, as it enables resources to include elements related to interest pixels in the memory. Unfortunately, in the *Working Memory* scene object patches whose shape features do not correspond to those defined by the observer may appear at time instant t. This is precisely because their shape characteristics have not yet been obtained. But, if these spots shape features really do not seem to be interesting for the observer, they will appear as "inhibited" in $t+1$ in the *Interest Map* (now, in $t+1$ their shape features will have been obtained). And, this means that in $t+1$ they will disappear from the *Working Memory*. Thus, the *Working Memory* has to be considered as a noisy memory. Scene object spots appear and disappear at each input image frame, as they fulfill or do not fulfill the desired spot shape features. In the same way that we have gotten sensitivity, we need some mechanism to endow the system with stability.

In order to provide stability to the system, that is to say, in order to obtain at each frame only objects with the desired features, we have to provide *Attention Reinforcement* by means of accumulative mechanism followed by a threshold. Accumulation is performed on pixels that have a value different from 0 (pixels that do not belong to labeled zones) in the *Working Memory*. The result of this process offers as output the *Attention Focus*, $AF[x,y,t]$.

Moreover, to obtain the *Attention Focus*, an intermediate memory called *Attention Map*, $AM[x,y,t]$, is used. In particular, pixels that appear with a value different from 0 in the *Working Memory* reinforce attention in the *Attention Map*, whilst those that appear as a 0 decrement the attention value. This accumulative effect followed by a threshold maintains "active" a set of pixels that belong to a group of scene object of interest to the observer. Hence, this is a charge/discharge process similar to the one explained in motion detection:

$$AM[x,y,t] = \begin{cases} \max(AM[x,y,t-1] - D_{AM}, Ch_{\min}) \\ \quad \text{if } WM[x,y,t] = 0 \\ \min(AM[x,y,t-1] + C_{AM}, Ch_{\max}) \\ \quad \text{if } WM[x,y,t] > 0 \end{cases} \quad (14)$$

where D_{AM} and C_{AM} are the attention-map-discharge constant and the attention-map-recharge constant, respectively. Now, based on the information provided by the

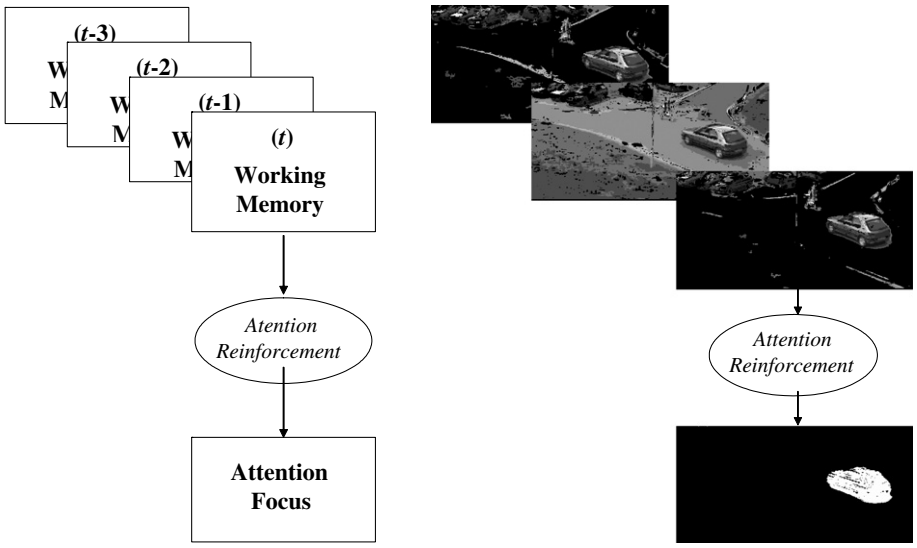

Fig. 13. Subtask "Attention Reinforcement".

Attention Map, objects need to be labeled in the *Attention Focus*. This is performed using an initial value at each pixel of *the Attention Focus* as shown in the following equation:

$$v[x,y] = \begin{cases} (x * NC + y) + 1 & \text{if } AM[x,y,t] > \theta \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

This initial value is contrasted with the values of the neighbors until a common value for all pixels of a same moving object is reached:

$$v[x,y] = \begin{cases} 0 & \text{if } v[x,y] = \min(v[\alpha,\beta]) = 0 \\ \min(v[\alpha,\beta]) & \text{if } 0 < \min(v[\alpha,\beta]) < v[x,y] \\ v[x,y] & \text{if } 0 < v[x,y] < \min(v[\alpha,\beta]) \end{cases} \quad \forall [\alpha,\beta] \in [x \pm 1, y \pm 1] | 0 < v[\alpha,\beta] \quad (16)$$

Finally, the value agreed is assigned to the *Attention Focus*:

$$AF[x,y,t] = v[x,y] \quad (17)$$

Fig. 13 shows the result of the accumulative computation on the *Attention Focus* and the later threshold. In this figure, pixels drawn in white color on black background represent image elements where attention has been focused.

4. Data and results

To test the performance of the proposed model, two image sets have been used. These sets are thought to show the versatility of our *Dynamic Visual Attention* model in object segmentation in indefinite sequences of video images. Therefore, we show the results of applying our model to scenes captured by a still and a moving camera. Notice that the method is dependent on the specific chosen scenario, in the sense that the parameters have to be tuned for the scenario and for each class of object to pay attention on. Fortunately, this parameter tuning does not depend on each different situation stored in a video sequence taken from the camera, but only on the predefined attention focuses. And this operation has only to be performed once.

The static scene, composed of 42 image frames of size 256×256 pixels, is the famous "Traffic intersection sequence" recorded at the Ettlinger-Tor in Karlsruhe by a stationary camera, copyright © 1998 by H.-H. Nagel, downloaded from the Institut für Algorithmen und Kognitive Systeme (H.-H. Nagel und Mitarbeiter) web pages.

Now, to illustrate the usefulness of the model with a moving camera a scene called "Horses" belonging to the movie "The Living Daylights" copyright © 1987 by MGM/UA has been used. This scene is composed of 80 frames of 122×512 pixels.

4.1. Results from static camera

This scene introduces some moving vehicles. Our aim is initially to capture attention on all moving objects of car class, independently of their velocities. That is to say, attention is focused on vehicles of an upper limited size. This way the bus present in the scene should not be classified as belonging to the attention focus. From frame 6 on, attention will be fixed on only one of the cars, performing this way a tracking task. Thus, the selection mode is changed at that frame.

First, parameters to detect and results of selecting the cars in motion in the scene as the attention focus are shown. In this case, the overlap has been $S = 8$. Table 1 shows the parameters used (as well as their values) to get

Table 1
Spot shape features used in *Working Memory*

Parameter	Value (number of pixels)
Spot maximum size: s_WM_{max}	90
Spot maximum width: w_WM_{max}	30
Spot maximum height: h_WM_{max}	30

Table 2
Object shape features used in *Attention Focus*

Parameter	Value (in pixels)	Value (ratios)
Object minimum size: s_AF_{min}	50	
Object maximum size: s_AF_{max}	200	
Object minimum width: w_AF_{min}	5	
Object maximum width: w_AF_{max}	80	
Object minimum height: h_AF_{min}	5	
Object maximum height: h_AF_{max}	50	
Object minimum width–height ratio: hw_AF_{min}		0.1
Object maximum width–height ratio: hw_AF_{max}		5
Object minimum compactness: c_AF_{min}		0.3
Object maximum compactness: c_AF_{max}		1.0

Table 3
Parameters of the *Attention Map*

Parameter	Values
Charge constant: C_{AM}	50
Discharge constant: D_{AM}	250
Threshold: θ	200

the patches' shapes in the *Working Memory*. Similarly, in Table 2 we show the parameters and values for the object's shapes in the *Attention Focus*. Lastly, the parameters used to calculate the *Attention Map* are offered in Table 3. Results are shown in Fig. 14.

In this figure you may notice some images of the sequence of selective attention on moving cars in different time instants. In row (a) some input images of the "Ettlinger-Tor" sequence are shown, namely at time instants $t = 3$, $t = 5$, $t = 6$, and $t = 42$. As it can be observed in Fig. 14, in row (b) "Active" pixels in the *Interest Map*, pixels where motion has been detected between two consecutive time instants, are shown. This is the result of calculating the presence of motion in the example. Remember that, in the output of this subtask, a pixel drawn in white color means that there has been variation in the grey-level band of the pixel in instant t with respect to the previous instant $t - 1$. There are pixels belonging to the desired objects, as well as to other parts of the image due to some variations in illumination in the scene. In the same figure, we have drawn in black color the "inhibited" pixels as well as the "neutral" pixels. In column (c) you can see the contents of the *Working Memory*, and in column (d) the *Active Attention Focus*. Lastly, on column (e) the *Active Attention Focus* has been overlapped with the input image. Fig. 14d shows the result of the accumulative computation on the *Active Attention Focus* and the later threshold. In this figure, pixels drawn in white color on black background represent image elements where attention has been focused and reinforced through time.

The silhouette of the tracked car (new mode) corresponds to the pixels of interest at $t = 6$. The *Working Memory* at $t = 6$ is composed of all the elements that include pixels of interest. So, in this particular case, part of the road appears apart from the tracked vehicle. As it can be observed, in the *Attention Focus* at $t = 6$ only the selected car appears, as the attention has been focused on this precise object.

In this example, we may notice that the attention focus really corresponds to the observer's intentions. First, attention is paid on all the moving cars. This example is very helpful to highlight some pros and cons of the described method. First, it is able to discriminate moving objects in an indefinite sequence into different classes of objects. This has been shown by the elimination of the bus in the scene through shape features parameterization. But some problems related to working with grey-level bands affect our method. This is why the bus is incorrectly decomposed into a priori smaller cars.

Table 4 shows the time spent to process each frame in our simulation environment, a Pentium IV personal computer running at 2.4 GHz and with a 512 MB memory under operating system Windows XP. The simulation has been programmed under Visual C++ version 6.0. Image size is a very important factor in our system, as well as working mode, as you may note on Table 4. To be competitive and to be able to process in real time the images cannot be too large or we have to consider the possibility to compute in specialized hardware.

It is important to notice that it is necessary to process a minimum number of frames to obtain the attention focus. This number of frames is related to parameters C_{AM}, the attention charge constant, and θ, the attention threshold. The value of θ is always greater than C_{AM} to enable the analysis of the shape of the elements of the *Working Memory*. Thus, it is necessary to process at least two transitions, that is to say, three frames, to get the *Attention Focus*.

4.2. Results from moving camera

In this second example, the camera that captures the scene is continuously performing a translation motion to follow the movement of the horses. Again two different applications have been done. The aim of the first one is to get all the moving horses on the scene as the focus of attention. The second application's objective is to track one determined element.

Next parameters (Tables 5–8) and results (Fig. 15) are shown when all the horses are the attention focus of the system. When changing the working mode to restrict to the tracking of one single element as determined by the attention focus, we get the results offered at Fig. 16.

Again, Table 8 shows the time spent to process each frame in our simulation environment. To highlight that the size of the objects paid attention on is also relevant

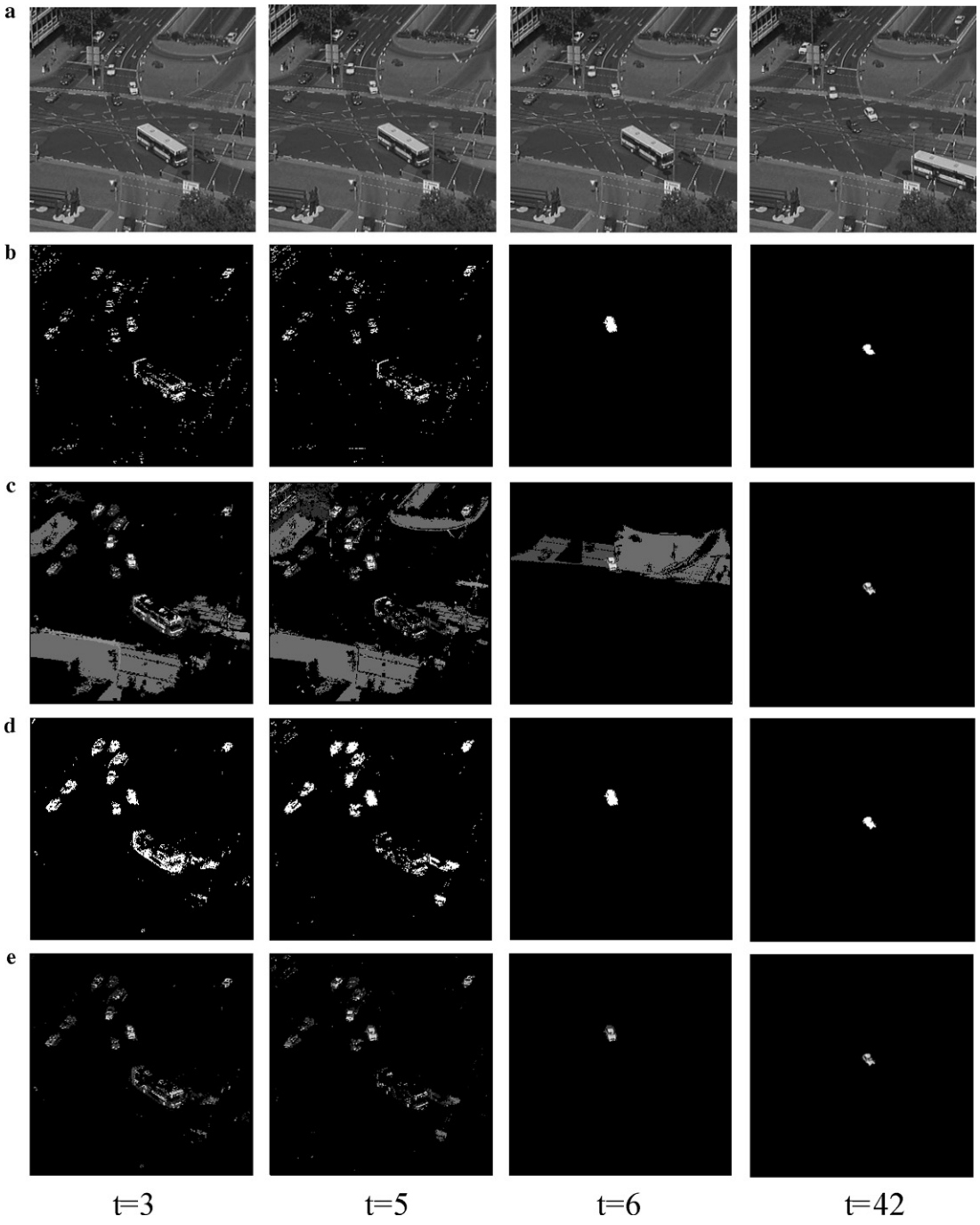

Fig. 14. Sequence of selective attention on car(s) in different time instants. (a) Input image. (b) "Active" pixels of the Interest Map. (c) Working Memory. (d) Attention Focus. (e) Attention Focus overlapped with input image.

Table 4
Performance

Size (in pixels)	Mode	Time/frame (in s)
256×256	Attention to multiple objects	2.2
256×256	Object tracking	0.6
128×128	Attention to multiple objects	0.25
128×128	Object tracking	0.08

Table 5
Spot shape features used in *Working Memory*

Parameter	Value (in number of pixels)
Spot maximum size: $s\_WM_{\max}$	500
Spot maximum width: $w\_WM_{\max}$	40
Spot maximum height: $h\_WM_{\max}$	40

Table 6
Object shape features used in *Attention Focus*

Parameter	Value (in pixels)	Value (ratios)
Object minimum size: $s\_AF_{\min}$	300	
Object maximum size: $s\_AF_{\max}$	4000	
Object minimum width: $w\_AF_{\min}$	20	
Object maximum width: $w\_AF_{\max}$	110	
Object minimum height: $h\_AF_{\min}$	−20	
Object maximum height: $h\_AF_{\max}$	100	
Object minimum width–height ratio: $hw\_AF_{\min}$		0.4
Object maximum width–height ratio: $hw\_AF_{\max}$		1.3
Object minimum compactness: $c\_AF_{\min}$		0.2
Object minimum compactness: $c\_AF_{\max}$		0.9

Table 7
Parameters of the *Attention Map*

Parameter	Values
Charge constant: C_{AM}	50
Discharge constant: D_{AM}	200
Threshold: θ	201

Table 8
Performance

Size (in pixels)	Mode	Time/frame (in s)
122×512	Attention to multiple objects	3.7
122×512	Object tracking	2.6
66×256	Attention to multiple objects	0.45
66×256	Object tracking	0.29

to the processing time. In comparison to the "Ettlinger-Tor" sequence, processing time of the "Horses" sequence is much higher. This is precisely due to the size of the objects tracked.

5. Conclusions

A model of dynamic visual attention capable of segmenting and tracking objects in a real scene has been introduced in this paper. The model enables to focus the attention in every moment on objects that possess certain features and to eliminate objects that are not of interest. The features used are related to the color, motion and shape of the elements present in the dynamic scene. The model may be used to monitor real environments indefinitely in time. Elements are considered to be of interest depending on the observer's commands. That is to say, a same scene may obtain different elements of interest just by changing the intentions.

On the opposite to computational models based on the space (spotlight, zoom), where attention is paid on one zone of the image, this paper proposes an object-based computational model, which enables to capture attention on one or various objects of the image. One of the mostly referenced selective attention models based on the spotlight metaphor is the Koch and Ullman model [25]. Its main advantage is that the model is biologically plausible.

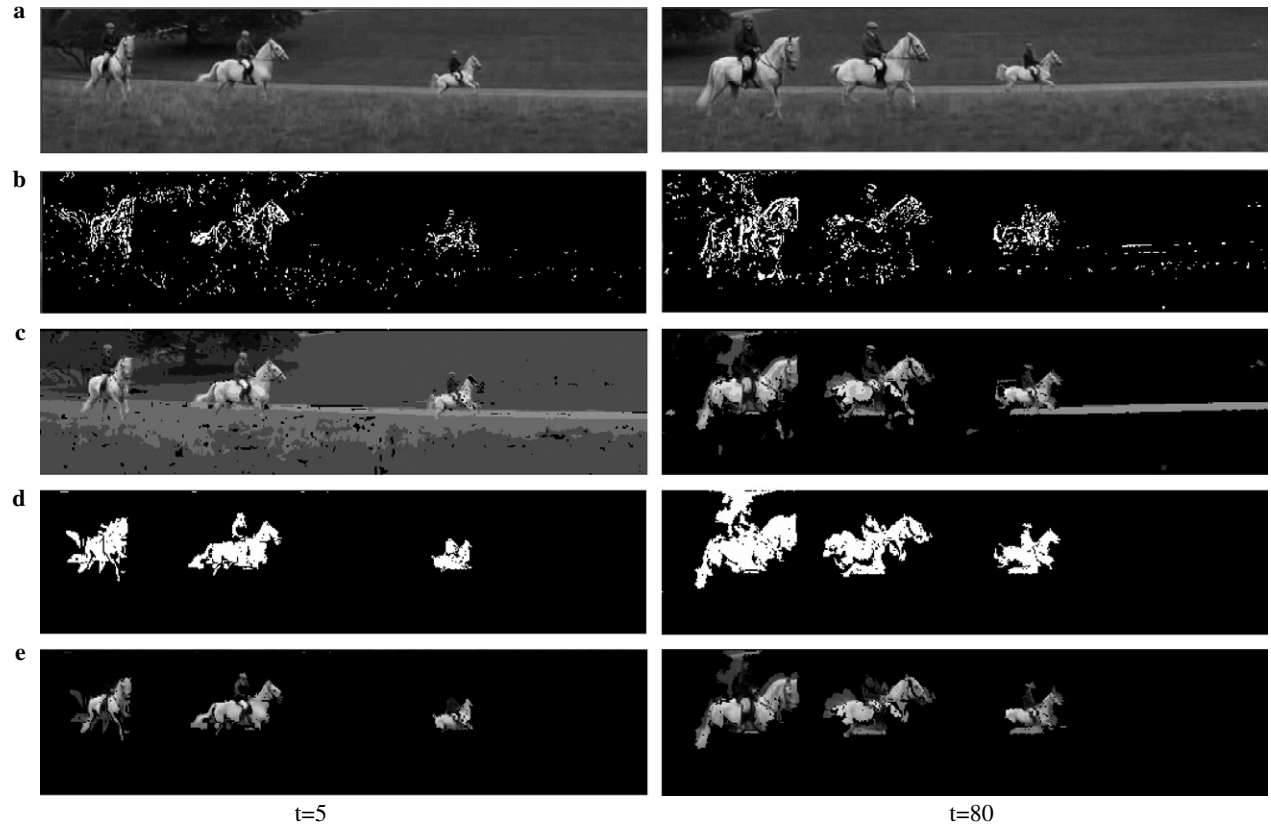

Fig. 15. Sequence of selective attention on all horses in different time instants. (a) Input image. (b) "Active" pixels of the Interest Map. (c) Working Memory. (d) Attention Focus. (e) Attention Focus overlapped with input image.

Fig. 16. Sequence of selective attention on one horse in different time instants. (a) Input image. (b) "Active" pixels of the Interest Map. (c) Working Memory. (d) Attention Focus. (e) Attention Focus overlapped with input image.

Unfortunately, its disadvantage is that it is restricted to static images. In dynamic environments the model of Backer and Mertsching [1] is of great interest to us. Indeed, the model proposed in our paper offers some important conceptual similarities with the recently described model by these authors [1], although our approach is quite different.

Similarly to [1], in our approach there is a feature extraction and integration step. In our model for each input image pixel grey-level band, motion presence, velocity and acceleration is extracted. For each element present in the *Working Memory* width, height and size features are calculated. And, for each object in the *Attention Focus* width, height, size, height-width relation and compactness features are obtained. All these features are integrated into the resulting *Interest Map*.

Again, there is a similarity in the use of selection phases as done by Backer and Mertsching. In our model the first selection is performed in the *Working Memory Generation* task, where elements that include "active" pixels of the *Interest Map* are selected. The second selection is obtained by means of the *Attention-Reinforcement* task. Accumulative computation on elements generated at each time instant in the so-called *Working Memory* is the result of this operation, incrementing this way the computational capacity of our approach.

Two examples to show the performance of our *Dynamic Visual Attention* model have been offered in this paper. And satisfactory results in scenes captured by a static camera as well as by a moving camera have been presented.

Acknowledgements

This work is supported in part by the Spanish CICYT TIN2004-07661-C02-01 and TIN2004-07661-C02-02 grants, and the Junta de Comunidades de Castilla-La Mancha PBI06-0099 grant. "Traffic intersection sequence at the Ettlinger-Tor in Karlsruhe", courtesy of Universität Karlsruhe, Fakultät für Informatik, Institut für Algorithmen und Kognitive Systeme, Kognitive Systeme (H.-H. Nagel und Mitarbeiter). The authors are also thankful for the permission to use the "PETS2001 Datasets, Dataset 1: Moving people and vehicles" available from The University of Reading, UK. "The Living Daylights", courtesy of MGM/UA. The authors are thankful to the anonymous reviewers for their very helpful comments.

References

[1] G. Backer, B. Mertsching, Two selection stages provide efficient object-based attentional control for dynamic vision, in: Proceedings of the International Workshop on Attention and Performance in Computer Vision, 2003, pp. 9–16.

[2] S. Baluja, D.A. Pomerleau, Expectation-based selective attention for visual monitoring and control of a robot vehicle, Robotics and Autonomous Systems 22 (3-4) (1997) 329–344.

[3] C. Balkenius, N. Hulth, Attention as selection-for-action: a scheme for active perception, in: Proceedings of EUROBOT'99. <www.lucs.lu.se/People/Christian.Balkenius/PostScript/Balkenius.Hulth.1999.pdf./>, 1999.
[4] J. Bonaiuto, L. Itti, The use of attention and spatial information for rapid facial recognition in video, Image and Vision Computing (2005), in press, doi:10.1016/j.imavis.2005.09.008.
[5] F. Caetano, J. Waldmann, Attentional management for multiple target tracking by a binocular vision head, SBA Controle & Automação 11 (2000) 187–204.
[6] D. Chung, R. Hirata, T.N. Mundhenk, J. Ng, R.J. Peters, E. Pichon, A. Tsui, T. Ventrice, D. Walther, P. Williams, L. Itti, A new robotics platform for neuromorphic vision: Beobots, in: Biologically Motivated Computer Vision, Springer, Germany, 2002, pp. 558–566.
[7] L. Czúni, T. Szirányi, Motion segmentation and tracking with edge relaxation and optimization using fully parallel methods in the cellular nonlinear network architecture, Real-Time Imaging 7 (2001) 77–95.
[8] G. Deco, J. Zihl, Top-down selective visual attention: a neurodynamical approach, Visual Cognition 8 (1) (2001) 119–140.
[9] R. Desimone, L.G. Ungerleider, Neural mechanisms of visual perception in monkeys, in: Handbook of Neuropsychology, Elsevier, Amsterdam, 1989, pp. 267–299.
[10] M.A. Fernández, J. Mira, Permanence memory: a system for real time motion analysis in image sequences, in: IAPR Workshop on Machine Vision Applications, 1992, pp. 249–252.
[11] M.A. Fernández, J. Mira, M.T. López, J.R. Álvarez, A. Manjarrés, S. Barro, Local accumulation of persistent activity at synaptic level: application to motion analysis, in: From Natural to Artificial Neural Computation, Springer, Germany, 1995, pp. 137–143.
[12] M.A. Fernández, A. Fernández-Caballero, M.T. López, J. Mira, Length-speed ratio (LSR) as a characteristic for moving elements real-time classification, Real-Time Imaging 9 (2003) 49–59.
[13] A. Fernández-Caballero, J. Mira, M.A. Fernández, M.T. López, Segmentation from motion of non-rigid objects by neuronal lateral interaction, Pattern Recognition Letters 22 (14) (2001) 1517–1524.
[14] A. Fernández-Caballero, J. Mira, A.E. Delgado, M.A. Fernández, Lateral interaction in accumulative computation: a model for motion detection, Neurocomputing 50 (2003) 341–364.
[15] A. Fernández-Caballero, M.A. Fernández, J. Mira, A.E. Delgado, Spatio-temporal shape building from image sequences using lateral interaction in accumulative computation, Pattern Recognition 36 (5) (2003) 1131–1142.
[16] A. Fernández-Caballero, J. Mira, M.A. Férnandez, A.E. Delgado, On motion detection through a multi-layer neural network architecture, Neural Networks 16 (2) (2003) 205–222.
[17] R.C. Gonzalez, R.E. Woods, Digital Image Processing, second ed., Addison-Wesley, Reading, MA, 2002.
[18] N. Götze, B. Mertsching, S. Schmalz, S. Drüe, Multistage recognition of complex objects with the active vision system NAVIS, in: Aktives Sehen in Technischen und Biologischen Systemen, 1996, pp. 186–193.
[19] O. Hasegawa, K. Yokosawa, M. Ishizuka, Real time parallel and cooperative recognition of human face for a naturalistic visual human interface, Systems and Computers 25 (11) (1994) 11–23.
[20] D. Heinke, G.W. Humphreys, Selective Attention for Identification Model: simulating visual neglect, Computer Vision and Image Understanding 100 (2005) 172–197.
[21] D. Heinke, G.W. Humphreys, G. diVirgilo, Modeling visual search experiments: Selective Attention for Identification Model (SAIM), Neurocomputing 44 (2002) 817–822.
[22] R. Herpers, K. Derpanis, W.J. MacLean, G. Verghese, M. Jenkin, E. Milios, A. Jepson, J.K. Tsotsos, SAVI: An actively controlled teleconferencing system, Image and Vision Computing 19 (2001) 793–804.
[23] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1254–1259.
[24] A.K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, USA, 1989.
[25] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurobiology 4 (1985) 219–227.
[26] M.T. López, M.A. Fernández, A. Fernández-Caballero, A.E. Delgado, Neurally inspired mechanisms for the dynamic visual attention map generation task, in: Computational Methods in Modeling Computation, Springer, Germany, 2003, pp. 694–701.
[27] A. Maki, P. Nordlund, J.-O. Eklundh, Attentional scene segmentation: integrating depth and motion, Computer Vision and Image Understanding 78 (3) (2000) 351–373.
[28] B. Mertsching, M. Bollmann, A. Massad, S. Schmalz, Recognition of complex objects with an active vision system, in: Proceedings of the International ICSC/IFAC Symposium on Neural Computation, ICSC Academic Press, 1998, pp. 469–475.
[29] T. Minato, M. Asada, Image feature generation by visio-motor map learning towards selective attention, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2001, pp. 1422–1427.
[30] T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding 81 (2001) 231–268.
[31] M. Mozer, The Perception of Multiple Objects: A Connectionist Approach, MIT Press, Cambridge, MA, 1991.
[32] L. Paletta, A. Pinz, Active object recognition by view integration and reinforcement learning, Robotics and Autonomous Systems 31 (1–2) (2000) 71–86.
[33] M.I. Posner, M.E. Raichle, Images of Mind, Scientific American Library, New York, 1994.
[34] E.O. Postma, H.J. van den Herik, P.T.W. Hudson, SCAN: a scalable model of attentional selection, Neural Networks 10 (6) (1997) 993–1015.
[35] A.L. Rothenstein, J.K. Tsotsos, Attention links sensing to recognition, Image and Vision Computing (2006), in press, doi:10.1016/j.imavis.2005.08.011.
[36] A.A. Salah, E. Alpaydin, L. Akarun, A selective attention-based method for visual pattern recognition with applications to handwritten digit recognition and face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (3) (2002) 420–425.
[37] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psychology 12 (1980) 97–136.
[38] S.P. Vecera, Toward a biased competition account of object-based segregation and attention, in: Brain and Mind, Kluwer Academic Publishers, The Netherlands, 2000, pp. 353–384.
[39] T. Wada, T. Matsuyama, Multiobject behavior recognition by event driven selective attention method, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 873–887.
[40] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, C. Koch, Attentional selection for object recognition: a gentle way, in: Biologically Motivated Computer Vision, Springer, Germany, 2002, pp. 472–479.
[41] J.M. Wolfe, Guided Search 2.0. A revised model of visual search, Psychonomic Bulletin and Review 1 (1994) 202–238.
[42] Y. Ye, J.K. Tsotsos, Sensor planning for 3D object search, Computer Vision and Image Understanding 73 (2) (1999) 145–168.
[43] K.J. Yonn, I.S. Kweon, Moving object segmentation algorithm for human-like vision system, in: 1st International Workshop on Human-Friendly Welfare Robotic Systems, 2000, pp. 109–114.
[44] D.S. Zhang, G. Lu, Segmentation of moving objects in image sequence: a review, Circuits, Systems and Signal Processing (Special Issue on Multimedia Communication Services) 20 (2) (2001) 143–183.