Multi-feature Bottom-up Processing and Top-down Selection for an Object-based Visual Attention Model

Antonio J. Palomino, Rebeca Marfil, Juan P. Bandera, and Antonio Bandera

Universidad de Málaga,
Departamento de Tecnología Electrónica, ETSI Telecomunicación,
Campus Universitario de Teatinos, 29071, Málaga, Spain
ajpalomino@uma.es

Abstract. Artificial vision systems can not process all the information that they receive from the world in real time because it is highly expensive and inefficient in terms of computational cost. However, inspired by biological perception systems, it is possible to develop an artificial attention model able to select only the relevant part of the scene, as human vision does. This paper presents an attention model which draws attention over perceptual units of visual information, called proto-objects, and which uses a linear combination of multiple low-level features (such as colour, symmetry or shape) in order to calculate the saliency of each of them. But not only bottom-up processing is addressed, the proposed model also deals with the top-down component of attention. It is shown how a high-level task can modulate the global saliency computation, modifying the weights involved in the basic features linear combination.

Keywords: visual attention, top-down selection, action-perception loop

1 Introduction

Human vision system presents an interesting set of features of adaptability and robustness that allows it to analyse and process the visual information of a complex scene in a very efficient manner. Research in Psychology and Physiology demonstrates that the efficiency of natural vision has foundations in *visual attention*, which is a process that filters out irrelevant information and limits processing to items that are relevant to the present task [1].

In the past few years, emphasis has increased in the development of robot vision systems that are inspired by the model of natural vision. An artificial attention system allows to optimize the required computational resources due to they can be focussed on the processing of a set of selected regions, which are important for the current task, instead of the whole image.

This ability is specially useful when developing a social robot, that is, an embodied agent which is part of a heterogeneous community of people and other robots [2]. In this case, added to the increased efficiency mentioned above, the

agent is able to process the visual information in the same way that people do. Thus, the interaction between a human and a robot becomes easier because both of them share the representation of their surrounding world.

According to psychological studies [3–5], there exist two contributions in the computation of how relevant an object is. On the one hand, any object has a saliency value by itself. Thus, an object is more relevant than other depending on its specific basic features (colour, shape, symmetry, location...). This is the so-called *bottom-up* component in attention. On the other hand, the ongoing task also imposes a particular relevance for each object. This contribution is known as the *top-down* part of attention. The final saliency value of an object in a scene is a combination of both contributions.

This paper introduces an attention model able to obtain the different visual entities in an image and, then, compute the most salient parts taking into account different simple features (colour and intensity contrasts, symmetry, orientation, roundness, proximity and dominant colours). The model also modifies the influence of each feature in the saliency computation in order to draw attention to those objects which are relevant to the current executed task. A previous version of the system was presented in [6]. That model used only 4 features based on colour and location to compute saliency so it lacks characterizing the objects in the scene. The model presented here is enhanced with 7 new features that include shape information of objects. Experimental results reveal that the new approach is more efficient guiding the top-down component of attention and provides a richer description of the elements in the scene.

The remainder of this paper summarizes the biological foundations and present work on artificial attention systems in section 2. In sections 3, 4 and 5, the different parts of the proposed approach are described. Experimental results are shown in section 6. Finally, conclusions are presented in section 7.

2 Related Work

From the psychological point of view, the development of artificial visual attention systems is mainly based on the so-called *early-selection* theories. These theories postulate that the selection of a relevant region precedes pattern recognition. Therefore, attention is drawn by simple features (such as colour, location, shape or size) and attended entities do not have full perceptive meaning, i.e., maybe they do not correspond to real objects.

Two complementary theories are the most influential ones regarding artificial attention systems: Treisman's *Feature Integration Theory* [3] and Wolfe's *Guided Search* [4, 5].

The first one suggests that the human vision system detects separable features in parallel in an early step of the attention process. According to this model, methods compute image features in a number of parallel channels in a *pre-attentive* task-independent stage. Then, the extracted features are integrated through a *bottom-up* process into a single saliency map which codes the relevance of each image entity.

Several years later, Wolfe proposed that a *top-down* component in attention can increase the speed of the process giving more relevance to those parts of the image corresponding to the current task. These two approaches are not mutually exclusive and, nowadays, some efforts in computational attention are being conducted to develop models which combine a bottom-up processing stage with a top-down selection [7] process.

Furthermore, attention theories introduce another important concept: the *Inhibition of Return*. This mechanism implies that an already attended entity should not be selected again for some time. Otherwise, the most relevant element would always be selected. This concept also applies in dynamic environments: the attended element remains the same despite the movement. That is, an element which changes its location is not considered as a new one if it does not disappear from the image.

Finally, Psychophysics studies also refer to how many elements can be attended at the same time. Bundesen establishes in his *Theory of Visual Attention* [8] that there exists a short-term memory where recently attended elements are stored. This memory has a fixed capacity usually reduced up to 3 or 5 elements.

The first artificial attention models mainly followed the guidelines established by Treisman's theory. For example, in the models proposed by Itti and Koch [9, 10] the saliency of each pixel is computed based on a set of basic features. They were pure bottom-up, static models. Later, Navalpakkam and Itti [7] modified Itti's original model in order to add a multi-scale object representation in a long-term memory. The multi-scale object's features stored in this memory determine the relevance of the scene features depending on the current executed task, implementing, therefore, a top-down behaviour. While these methods compute saliency pixel by pixel, Aziz's approach [11] proposes a *region-based* model that performs a pixel clustering prior to the saliency computation.

As an alternative to the previous *space-based* models, where attention deploys on an unstructured region of the scene rather than on an object, *object-based* models of visual attention provide a more efficient visual search. These models are based on the assumption that the boundaries of segmented objects, and not just spatial position, determine what is selected and how attention is drawn [12]. Therefore, these models reflect the fact that perception abilities must be optimized to interact with objects and not just with disembodied spatial locations. For example, Sun and Fisher [13] present a grouping-based saliency method and a hierarchical selection of attention at different perceptual levels (points, regions or objects). The problem of this model is that the groups are manually drawn. Orabona et al. [14] propose a model of visual attention based on the concept of *proto-objects* [15] as units of visual information that can be bound into a coherent and stable object. They compute these proto-objects by employing the watershed transform to segment the input image using edge and colour features in a pre-attentive stage. The saliency of each proto-object is computed taking into account top-down information about the object to perform a task-driven search. Yu et al. [16] propose a model of attention that segments the scene into proto-objects in a bottom-up strategy based on Gestalt theories. After that, in a

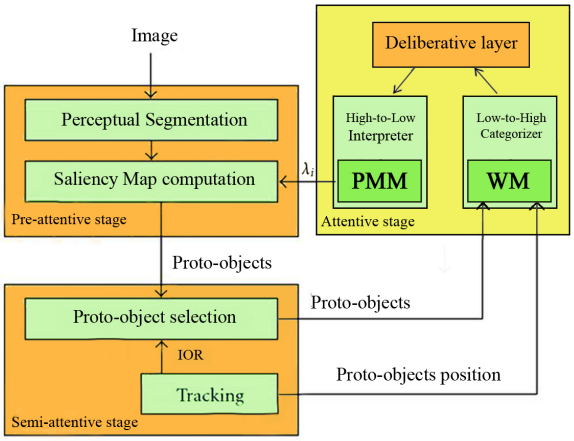


Fig. 1: Overview of the Object-Based Attention Model

top-down way, the saliency of the proto-objects is computed taking into account the current task to accomplish by using models of objects which are relevant to this task. These models are stored in a long-term memory.

In addition to the models mentioned above, other approaches uses alternative paradigms to build artificial visual attention systems. For example, Judd et al. [17] propose to learn a model of saliency directly from human eye movement data, acquired using an eye-tracking system. Following a different philosophy, Tsotsos et al. [18] model visual attention by the selective tuning of complex neural networks. A complete review about recent attention models can be found in [19, 20].

3 Overview of the proposed model

The attention model presented in this paper is an extension of the one previously introduced in [6], a visual attention system for a social robot which works in a dynamic scenario. Fig. 1 shows an overview of the model.

The proposed attention system integrates task-independent bottom-up processing and task-dependent top-down selection. The units of attention are the so-called proto-objects [15]. These proto-objects are defined as the blobs of uniform colour and disparity of the image which are bounded by the edges obtained using a Canny detector. On the one hand, the bottom-up component determines the set of proto-objects in the image, describing them by a set of low-level features that are considered relevant to determine their corresponding saliency values. On the other hand, the top-down component weights the low-level features that characterize each proto-object to obtain a single saliency value depending on the task to perform.

In the pre-attentive stage, the different proto-objects in the image are extracted using a perceptual segmentation algorithm based on a hierarchical framework [21]. As the process to group image pixels into higher-level structures can be computationally complex, perceptual segmentation approaches typically combine a pre-segmentation step with a subsequent perceptual grouping step [22]. The pre-segmentation step performs the low-level definition of segmentation as the process of grouping pixels into homogeneous clusters and the perceptual grouping step conducts a domain-independent grouping which is mainly based on properties such as proximity, closure or continuity.

Then, the relevance of each proto-object is computed taking into account different low-level features weighted by a set of perception parameters (λ_i) stored in a *Perception-Modulation Memory* (PMM). From a psychological point of view, these perception parameters are closely related to the "attentional sets" proposed by Corbetta et al. [23]. While our previous approach computed only 4 low-level characteristics, the present model is able to handle up to 11 basic features in order to obtain the final saliency value. These features and saliency computation are deeply described in section 4.

The semi-attentive stage, deals with the management of the *Working Memory* (WM) and the Inhibition of Return (IOR). The WM establishes the maximum number of attended elements that can be maintained at once. It is a short-term memory where the system stores the recently attended objects and it has a reduced capacity, up to 5 elements [8]. Each proto-object in the WM is characterized by a set of descriptors including its saliency value, its position in the image, some basic properties, the different low-level features values and a time-to-live value which establishes the maximum time that the proto-object can stay in the WM. The saliency of a proto-object also depends on this last parameter, so the longer an element is kept in the WM, the lower its saliency is. A new proto-object get into the WM if and only if it has bigger saliency than the currently stored elements. If the memory is full, the least salient element is dropped out.

Regarding the IOR, it is typically implemented using a 2D inhibition map which contains suppression factors for one or more focusses of attention recently attended. This approach is valid to manage static scenarios, but it is not able to handle dynamic environments where inhibited proto-objects or the vision system itself are in motion. In our proposal, a tracker module keeps permanently updated the position of each element in the WM, allowing to manage not only moving objects but also camera and robot movements. Thereby, it is avoided to attend an already selected proto-object even if the proto-object changes its location in the image. Specifically, the tracker is based on the Comaniciu's mean-shift approach [24], a method which allows to track non-uniform colour regions in an image. When a proto-object is lost, it is also removed from the WM so the proposed attention model deals mainly with overt attention.

4 Multi-feature saliency computation

As it was aforementioned, the relevance of a proto-object is obtained as the combination of multiple basic features. In comparison to [6], a greater number of features is employed allowing a better characterisation of the properties of a proto-object, including information about shape, colour and localization. In order to have an homogenized calculus, all features values are normalized in the range $[0 \dots 255]$. In terms of computational cost, features are computed in parallel so adding more of them does not produce an important overhead in processing.

4.1 Colour contrast and Intensity contrast features

These features measure how different a proto-object is with respect to its surrounding in terms of colour and luminosity.

Since proto-objects are the result of a perceptual segmentation process, we can compute the colour contrast, $(ColCON)$, of a specific proto-object, $\mathcal{P}_i$, as the mean colour gradient along its boundary to the neighbours:

$$ColCON_i = \frac{S_i}{b_i} \sum_{j \in N_i} b_{ij} \cdot d\left(< C_i >, < C_j >\right) \tag{1}$$

where b_i is the perimeter of $\mathcal{P}_i$, N_i is the set of proto-objects which are neighbours of $\mathcal{P}_i$, b_{ij} is the length of the perimeter of $\mathcal{P}_i$ in contact with proto-object $\mathcal{P}_j$, $d\left[< C_i >, < C_j >\right]$ is the HSV colour distance between the colour mean values $< C >$ of proto-objects $\mathcal{P}_i$ and $\mathcal{P}_j$ and S_i is the mean saturation value of proto-object $\mathcal{P}_i$. Because of the use of S_i in the colour contrast equation, white, black and pure gray proto-objects are suppressed. To solve that, we compute the intensity contrast, $(IntCON)$, of a proto-object, $\mathcal{P}_i$, as the mean luminosity gradient along its boundary to the neighbours:

$$IntCON_i = \frac{1}{b_i} \sum_{j \in N_i} b_{ij} \cdot d\left(< I_i >, < I_j >\right) \tag{2}$$

being $< I_i >$ the mean luminosity value of the proto-object $\mathcal{P}_i$.

4.2 Proximity feature

Another important parameter in order to characterize a proto-object is to determine how near it is from the vision system location. Nowadays, not only stereo pairs of cameras but also cheaper sensors like Microsoft Kinect or ASUS Xtion provide accurate depth information of the captured image.

When using a sensor able to provides depth information directly (e.g. a RGBD camera or similar), the proximity, $(PROX)$, of a proto-object, $\mathcal{P}_i$, is directly obtained as the inverse of the depth value provided by the sensor:

$$PROX_i = \frac{1}{depth_i} \tag{3}$$

If we are using a stereo pair of cameras as depth sensor, the proximity can be obtained directly from disparity information.

Although pure depth information is not used in saliency computation, it is saved as a descriptor of the proto-object for further use.

4.3 Roundness feature

The next 3 features give us information about the shape of each proto-object. Roundness measurement reflects how similar to a circle a proto-object is. To calculate it, a traditional technique based on image moments is employed. Concretely, 3 different central moments are used:

$$\mu^i_{1,1} = \sum (x - \overline{x})(y - \overline{y}) \quad \forall (x, y) \in \mathcal{P}_i \tag{4}$$

$$\mu^i_{2,0} = \sum (x - \overline{x})^2 \quad \forall (x, y) \in \mathcal{P}_i \tag{5}$$

$$\mu^i_{0,2} = \sum (y - \overline{y})^2 \quad \forall (x, y) \in \mathcal{P}_i \tag{6}$$

being $(\overline{x}, \overline{y})$ the center of the proto-object $\mathcal{P}_i$.

The eccentricity (how different from a circle a region is) is computed directly from central moments:

$$ecc_i = \frac{(\mu^i_{2,0} - \mu^i_{0,2})^2 + 4{\mu^i_{1,1}}^2}{(\mu^i_{2,0} + \mu^i_{0,2})^2} \tag{7}$$

Finally, We compute the roundness (similarity to a circle), $(ROUND_i)$, for a proto-object, $\mathcal{P}_i$, as:

$$ROUND_i = 1 - ecc_i \tag{8}$$

4.4 Orientation feature

The orientation of a proto-object can also be obtained from central moments computed in (4), (5) and (6):

$$\varphi_i = \frac{1}{2} \arctan \left(\frac{2\mu^i_{1,1}}{\mu^i_{2,0} - \mu^i_{0,2}} \right) \tag{9}$$

The orientation of a proto-object, by itself, does not provide any useful information about its relevance. In fact, it is more interesting to compute saliency in terms of contrast with the orientation of other elements. The orientation contrast, $(OriCON)$, of a proto-object, $\mathcal{P}_i$, is obtained as:

$$OriCON_i = \sum_{j, j \neq i}^{\mathcal{P}} |\varphi_i - \varphi_j| \tag{10}$$

4.5 Symmetry feature

To compute the symmetry of a proto-object we use an approach similar to [11]. They propose a method to obtain symmetry using a scanning function $\psi(L, P_s)$ that counts the symmetric points around a point P_s along a line L. This procedure is repeated employing different lines of reference. For each line, the measure of symmetry is computed as:

$$S^{\theta} = \sum_{s=1}^{l} \frac{\psi(L, P_s)}{\alpha(R_i)} \tag{11}$$

where l and θ are the length and the angle of the line of reference and $\alpha(R_i)$ is the area of the region in order to normalize the result between 0 and 1.

Only an approximation of symmetry is needed in terms of attention systems. Thus, only 4 different angles for symmetry axes are considered: 0°, 45°, 90° and 135° respect to the orientation of the image (obtained in (9)).

In [11], the total measure of symmetry is computed as an average of the symmetry values in the different lines of reference. However, such strategy can define a region with only one axis of symmetry as asymmetric because non-symmetric axes cancel out the contribution of the symmetric one. As we are giving relevance to symmetry independently of the axis of symmetry, we compute the maximum symmetry, $(SYMM)$, for a proto-object, $\mathcal{P}_i$, as:

$$SYMM = max_{\theta}(S^{\theta}) \tag{12}$$

4.6 Dominant Colour features

Sometimes, an application requires objects of a specific colour to be more relevant than others regardless the number of them present in the scene. For example, a fireman robot may look for red extinguishers or a social robot is likely to search for people (who have a characteristic skin colour). For this reason, a set of dominant colour features are added to the attention system. Concretely, the proposed model includes saliency computation for 4 basic colours (red, blue, green and yellow) and for skin colour.

Regarding the similarity to basic colours, an HSV colour distance is employed. The hue and saturation values are compared with a reference. If the distance is less than a threshold Θ, the correspondent colour feature obtains a value of 255. Otherwise, the value is 0. The saturation value is used to avoid shadows to be marked as colours. The red, (RED), blue, (BLU), green, (GRN), and yellow, (TLW), correspondent colour for a proto-object, $\mathcal{P}_i$, are computed as:

$$RED_i = \begin{cases} 255 \text{ if } d\left(< C_i >, < C_{red} >\right) \leq \Theta_{red} \\ 0 \quad \text{otherwise} \end{cases} \tag{13}$$

$$BLU_i = \begin{cases} 255 \text{ if } d\left(< C_i >, < C_{blue} >\right) \leq \Theta_{blue} \\ 0 \quad \text{otherwise} \end{cases} \tag{14}$$

$$GRN_i = \begin{cases} 255 \text{ if } d\left(< C_i >, < C_{green} >\right) \leq \Theta_{green} \\ 0 \quad \text{otherwise} \end{cases} \tag{15}$$

$$YLW_i = \begin{cases} 255 \text{ if } d\left(< C_i >, < C_{yellow} >\right) \leq \Theta_{yellow} \\ 0 \quad \text{otherwise} \end{cases} \tag{16}$$

By definition, a proto-object can only belong to one (or none) of the basic colours.

Concerning the skin colour, the computation is based on the skin colour chrominance model proposed by Terrillon and Akamatsu [25]. Firstly, the image is transformed into the TSL colour space. Then, the Mahalanobis distance between the colour of the proto-object and the mean vector of the skin chrominance model is computed. If this distance is less than a threshold Θ_{skin}, the skin colour feature is marked with a value of 255. Otherwise, it is set to 0.

$$SKN_i = \begin{cases} 255 \text{ if } d_M\left(< C_i^{TSL} >, < C_{yellow}^{TSL} >\right) \leq \Theta_{skin} \\ 0 \quad \text{otherwise} \end{cases} \tag{17}$$

4.7 Saliency computation

The final saliency value, SAL_i, for each proto-object, $\mathcal{P}_i$, is obtained as a weighted sum of all the previous features:

$$sal_i = \boldsymbol{\lambda} \cdot \boldsymbol{f} \tag{18}$$

where $\boldsymbol{\lambda}$ is the weights vector, verifying $\sum_i \lambda_i = 1$, and $\boldsymbol{f}$ is the feature vector.

Expanding (18) in terms of (1), (2), (3), (8), (10), (12), (13), (14), (15), (16) and (17), we obtain the final expression for saliency:

$$\begin{aligned} SAL_i = {} & \lambda_1 \cdot ColCON + \lambda_2 \cdot IntCON + \lambda_3 \cdot PROX + \lambda_4 \cdot ROUND \\ & + \lambda_5 \cdot OriCON + \lambda_6 \cdot SYMM + \lambda_7 \cdot RED + \lambda_8 \cdot BLU \\ & + \lambda_9 \cdot GRN + \lambda_{10} \cdot YLW + \lambda_{11} \cdot SKN \end{aligned} \tag{19}$$

Eq. (19) shows that the more different to other proto-objects in the image, the more salient the proto-object.

5 Connecting Bottom-up processing and Top-down selection

The integration between bottom-up and top-down contributions is an important issue in visual attention systems. There are two main strategies to address this problem: *feature map fusion approaches* and *template-based approaches*. The first ones (e.g. [26]) extend Itti's model [9] in order to compute new feature maps based on high-level learned knowledge. Then, top-down maps are fused with the bottom-up ones and a final saliency master map is obtained. On the other hand, template-based models, such as the approach presented in [27], work with abstract templates of low-level features (colour, shape, symmetry, etc) of the target.

These models do not need any previous training of the system and can manage abstract information about the target ("look for green, rounded objects").

In an autonomous system, top-down information is usually provided by a deliberative layer (typically, a planner). Since the planning system defines what to do and, therefore, it must suggest what type of information is relevant or not to the attention module, it is interesting that the latter can deal with abstract predicates. Consequently, we propose a template-based approach which is more suitable to manage that kind of information.

In our proposal, both the Working Memory (WM) and Perception Modulation Memory (PMM) are the interface between early attention stages and the rest of the system, including the deliberative level. This interface includes a categorizer which is able to classify the perceived proto-objects into categories corresponding to high-level predicates. Furthermore, the PMM translates high-level instructions, splitting them into a set of low-level features to be highlighted in the scene, that is, a new vector of perception parameters, $\boldsymbol{\lambda}$. Therefore, it is allowed to change the way the vision system perceives the world in terms of a high-level decision.

The way the perception parameters are computed from the deliberative level is strongly dependent on each particular application. Therefore, there exists a particular *high-to-low* interpreter for each concrete problem. For example, a car-driving task looking for the nearest "stop" signal is translated to look for red and rounded proto-objects in the image which are close to the camera. Taking equation (19) as reference, the mentioned task implies large values for λ_3 (affecting proximity), λ_4 (affecting roundness) and λ_7 (affecting red dominant colour). On the contrary, the remainder perception parameters obtain a small value in comparison. When the task changes, a variation of perception parameters pops out different proto-objects in the scene.

6 Experimental results

The images have been obtained using a Microsoft Kinect sensor which provides both RGB image and depth information. The resolution of the images is 640x480 pixels. To process the information, a PC with an Intel Core2Duo processor at 2.66 GHz and 4 GB of DDR2 RAM at 800 MHz is employed. The software has been developed using *RoboComp*, an open-source robotics framework [28].

6.1 Features extraction

Fig. 2 shows the computation of the features described in section 4 from a real image taken in a lab. As it can be observed, the scene is compound by some coloured balls, a blue glass and a blue case in foreground and a person behind them. There is no special restriction about illumination or the elements in background. The result of perceptual segmentation in order to obtain the proto-objects in the scene is shown in fig. 2.b. It can be observed than most of the proto-objects in foreground correspond to real objects. However, it can

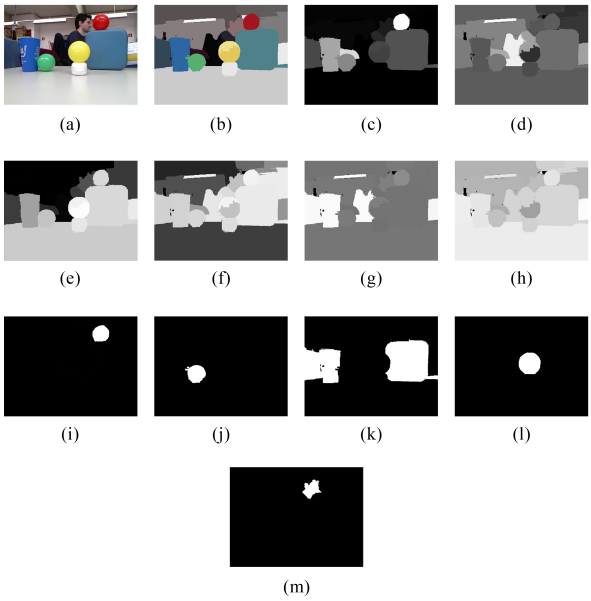


Fig. 2: Features maps. (a) original image; (b) perceptual segmentation (extraction of proto-objects); (c) colour contrast; (d) intensity contrast; (e) proximity; (f) roundness; (g) orientation contrast; (h) symmetry; (i) dominant colour: red; (j) dominant colour: green; (k) dominant colour: blue; (l) dominant colour: yellow; (m) skin colour.

happen that a real object is divided into two different proto-objects due to the segmentation process (e.g. the yellow ball in fig. 2.b).

6.2 Saliency computation and top-down selection

Since we compute saliency as a weighted sum of features, it is possible to change what elements in scene are going to be more relevant. Fig. 3 shows how it is possible to draw attention to different parts of the image changing the value of the perception parameters λ_i. For example, in fig. 3.c only roundness is taken as relevant feature so $\lambda_{ROUND} \gg \lambda_i, \forall i \neq ROUND$. It can be noted that not only balls but also squared objects (such as the blue case) have a high roundness. As it was mentioned in section 5, it is also possible to look for objects verifying 2 or more features. This can be observed, for example, in fig. 3.e, where the system is looking for green and rounded objects in scene. Thus, the proto-object corresponding to the green ball is the one with the highest saliency value.

In fig. 4, a search for round, blue and yellows objects ($\lambda_{ROUND} = \lambda_{BLU} = \lambda_{YLW} \gg \lambda_i, \forall i \neq ROUND \neq BLU \neq YLW$) is detailed. The 5 most salient proto-objects are shown in fig. 4.d. As it was expected, the object complying

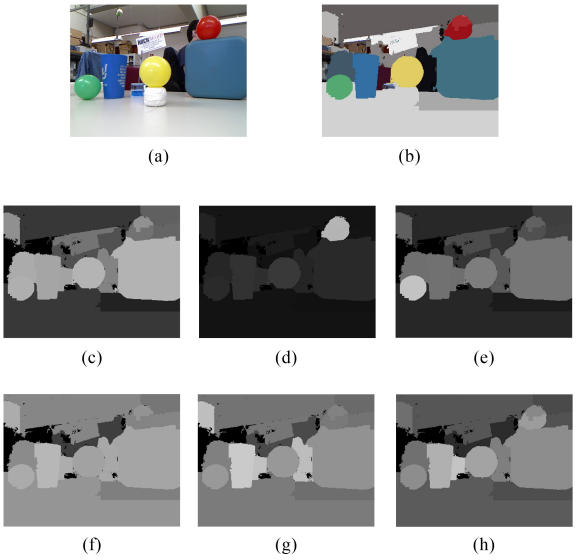
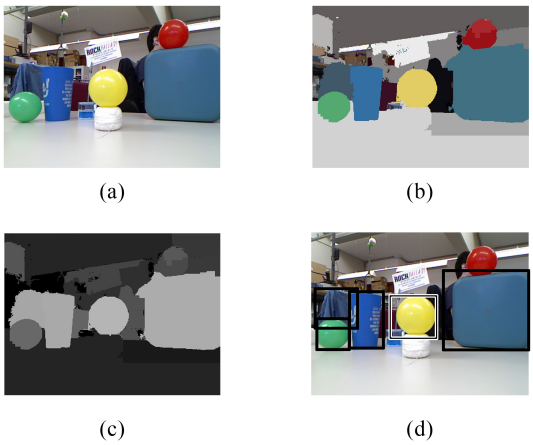


Fig. 3: Different saliency maps in terms of different perception parameters ($\boldsymbol{\lambda}$). (a) original image; (b) perceptual segmentation (extraction of proto-objects); (c) only roundness is relevance; (d) only red colour is relevant; (e) roundness and green colour are relevant; (f) only symmetry is relevant; (g) symmetry and orientation contrast are relevant; (h) symmetry and colour contrast are relevant.

with the most of the required features (the yellow ball) is selected as the most relevant. Although the blue case also verifies 2 features, the roundness of a square is a bit lower than the roundness of a ball. Thus, the blue case is selected as the second most relevant element. The remainder of the elements to be stored in the WM corresponds to blue or round objects.

7 Conclusions

In this paper we have presented an attention model integrating a bottom-up processing, based on simple features extraction, and a template-based top-down selection. The model deploys attention on proto-objects, units of visual information that can be bound into coherent and stable objects. The saliency of each proto-object is computed as a weighted sum of basic features describing colour, shape and location. The most relevant proto-objects are stored in a Working Memory able to update their position in a dynamic scenario. In order to connect the attention model with a deliberative layer, a Perception Modulation Memory is defined to store different values of the weights which define the saliency computation. Depending on the behaviour to achieve, the system can modify



(a) (b) (c) (d)

Fig. 4: Saliency computation looking for round, blue and yellow proto-objects. (a) original image; (b) perceptual segmentation (extraction of proto-objects); (c) saliency map; (e) proto-objects selected to be stored in the WM (the most salient one is marked with a black and white bounding-box).

the features taken into account to look for objects which are relevant for the ongoing task, changing the value of such weights.

Acknowledgements

This work has been partially granted by the Spanish Ministerio de Economía y Competitividad (MINECO) projects TIN2008-06196 and TIN2012-38079-C03-03. It has been also granted by Universidad de Málaga, International Campus of Excellence Andalucía Tech.

References

1. Duncan, J.: Selective attention and the organization of visual information. Journal of Experimental Psychology 113(4) (1984) 501-517
2. Dautenhahn, K., Billard, A.: Bringing up robots or the psychology of socially intelligent robots: From theory to implementation. In: Proc. of the third annual Conf. on Autonomous Agents, Seattle, Washington, United States (1999)
3. Treisman, A., Gelade, G.: A feature integration theory of attention. Cognitive Psychology 12(1) (1980) 97-136
4. Wolfe, J., Cave, K., Franzel, S.: Guided search: An alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human Perception and Performance 15(3) (1989) 419-433
5. Wolfe, J.: Guided search 2.0: a revised model of visual search. Psychonomic Bulletin and Review 1 (1994) 202-238

6. Palomino, A.J., Marfil, R., Bandera, J.P., Bandera, A.: A novel biologically inspired attention mechanism for a social robot. EURASIP Journal on Advances in Signal Processing (2011)
7. Navalpakkam, V., Itti, L.: Moelling the influence of task on attention. Vision Research 45(2) (2005) 205-231
8. Bundesen, C., Habekost, T., Kyllingsbaek, S.: A neural theory of visual attention and short-term memory (ntva). Neuropsychologia 49(6) (2011) 1446–1457
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1254-1259
10. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlaying neural circuitry. Human Neurobiology 4(4) (1985) 219-227
11. Aziz, Z., Mertsching, B.: Fast and Robust Generation of Feature Maps for Region-Based Visual Attention. IEEE Transactions on Image Processing 17(5) (2008) 633–644
12. Scholl, B.: Objects and attention: the state of art. Cognition 80(1-2) (2001) 1-46
13. Sun, Y., Fisher, R.: Object-based visual attention for computer vision. Artificial Intelligence 146(1) (2003) 77-123
14. Orabona, F., Metta, G., Sandini, G.: A proto-object based visual attention model. In Paletta, L., Rome, E., eds.: WAPCV 2007. LNCS (LNAI). Volume 4840., Heidelberg, Springer (2007) 198-215
15. Rensink, R.: Seeing, sensing, and scrutinizing. Vision research 40(10-12) (2000) 1469-1487
16. Yu, Y., Mann, K., Gosine, R.: An object-based visual attention model for robotic applications. IEEE Transactions on Systems, Man and Cybernetics B 40(3) (2010) 1-15
17. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: Computer Vision, 2009 IEEE 12th international conference on, IEEE (2009) 2106-2113
18. Tsotsos, J.K., Culhane, S.M., Kei Wai, W.Y., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artificial intelligence 78(1) (1995) 507-545
19. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM Transactions on Applied Perception (TAP) 7(1) (2010) 6
20. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(1) (2013) 185-207
21. Marfil, R., Molina-Tanco, L., Rodríguez, J., Sandoval, F.: Real-time object tracking using bounded irregular pyramids. Pattern Recognition Letters (28) (2007) 985-1001
22. Marfil, R., Bandera, A., Sandoval, F.: Comparison of perceptual grouping criteria within an integrated hierarchical framework. In Escolano, F., Torsello, A., eds.: Proceedings of the Graph-Based Representations in Pattern Recognition (GbRPR09). Volume 5534 of Lecture Notes in Computer Science., Venice, Italy, Springer (2009) 366-375
23. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nature reviews neuroscience 3(3) (2002) 201-215
24. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Transanction on Pattern Analysis and Machine Intelligence 25(5) (2003) 564-575
25. Terrillon, J., Akamatsu, S.: Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images. In: Proceedings of the 12th Conference on Vision Interface. (1999) 180-187

26. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 2049-2056
27. Tunnermann, J., Born, C., Mertsching, B.: Top-down visual attention with complex templates. In: International Conference on Computer Vision Theory and Applications. (February 2013) 370-377
28. Gutiérrez, M., Romero-Garcés, A., Bustos, P., Martínez, J.: Progress in robocomp. Journal of Physical Agents 7(1) (2013)