Target Search by Bottom-Up and Top-Down Fuzzy Information

EVELINA MARIA DE ALMEIDA NEVES[A], JOÃO EDUARDO BORELLI[B], ADILSON GONZAGA[C]

[a] Physics Institute at São Carlos, University of São Paulo, Brazil
evelinan@sel.eesc.sc.usp.br
[b,c] School of Engineering at São Carlos, Department of Electrical Engineering
University of São Paulo, São Carlos, 13560-250, SP, Brazil
[b]jborelli@peterpan.sel.eesc.sc.usp.br, [c]agonzaga@sc.usp.br

Abstract. One of the basic tasks assigned to the attentional mechanism is to decide which location in the visual field we must pay attention first. An object containing a distinctive feature can attract attention in a bottom-up way. By comparing one object with the others present in the scene, bottom-up conspicuity features are used to guide attention to the most different object. Top-down hints are based on the previous knowledge about the objects or on which features are important to locate them and also have a large influence on the attended locations. Inspired on the mechanisms of human visual attention we developed a new methodology to integrate bottom-up and top-down information by using fuzzy net containing three fuzzy subsystems. The first bottom-up subsystem allow us to combine features and infer with great flexibility some intuitive decision rules based on the visual perception principles such as the Gestalt laws. The second top-down subsystem combines different features according to the relevance of them in different tasks. Finally, the last subsystem integrates the information of the previous systems and gives a general salience index. The new methodology was tested in geometrical objects considering the features that attracts attention to human beings.

1 Introduction

One of the basic task assigned to the attentional mechanism is to decide which location in the visual field we must pay attention first. *Visual Search*, the ability to find one item in a visual world filled with other distracting items is not trivial and has been subject of research in the past 20 years [1-2]. The relative ease which humans can search for targets in a scene containing different objects, sharing different properties, is also the subject of this paper.

One of the great difficulty in these tasks is that, in order to distinguish the target from the distractors, a combination of features must be associated with a single object. Often called the *binding problem,* this requirement presents a serious hurdle to deal to when multiple objects are present. Psychophysical experiments suggest that people use covert visual attention to get around this problem [3]. In visual search two aspects of the problem are important: feature integration and localization. Feature integration is concerned with the interference between features of different objects when a parallel representation is used. Similarly, the interference between objects makes it difficult to recover the locations of individuals objects. Two main kinds of processes contribute in determining the value at each location: bottom-up and top-down. Objects that differ a lot from its environment, for example in color, size or orientation, attracts attention in a bottom-up way. In the top-down attention, the higher cognitive levels of the brain influence the attentional system to select in favor of a particular feature or of a combination of features. The first class is purely data-driven, whereas the second one includes constraints about the task and depends on the knowledge already gathered from the image.

At a variety of hierarchical levels (from the retina to the higher stages of the visual cortex), selection mechanisms discard most of the information in order to concentrate the limited processing resources on the most important and interesting parts of the visual input. To a higher level than just feature analysis, a form of selection would also be needed to ensure behavioral coherence (*attention for action*). Since visual perception is the mean that allows subjects to interact (manipulated, avoid, etc.) with the objects that compose their environment, a number of actions are continuously elicited and guided by object perception. Each of these actions requires the specification of a number of parameters that can be thought of as controlled by the identity, position and appearance of objects. Selective processing would then be necessary in order to isolate the information that defines the parameters for the appropriate actions. If each information processing stream, generated by some visual stimulus potentially leads to one action, there is the need to keep it separated from the others [4].

Johnston & Dark [5] defines attention as the differential processing of simultaneous sources of information. Visual Attention can be seen as a "glue"

0-7695-0878-2/00 $10.00 © 2000 IEEE

that ties internal (memory and knowledge) and external (objects and events) information sources. Without attention, characteristics cannot be related with other ones.

Computational work on visual attention has been influenced by theories proposed by cognitive neuroscience, in particular by the models suggested by Treisman's experiments on visual search [1-2]. The computational model proposed by Treisman distinguishes between a pre-attentive, massively parallel stage that process information about basic visual features (color, motion, depth cues, etc.) and a subsequent limited-capacity stage that performs other more complex operations, such as face recognition, reading and object identification and is applied over a limited portion of the visual field. The spatial deployment or the limited-capacity process is under attentional control.

Cave & Wolfe [5] defined another computational model inspired by Treisman's *feature integration theory*. In their model, called *Guided Search*, perception occurs in two stages: a parallel and a sequential one. In the parallel stage a set of feature maps is computed and each feature map evaluates how the feature in each location differs from the features of other locations (conspicuity measure). These values are then put together into a global map. High-level knowledge about the features of the target is used to weight the features in the merging step. In the sequential stage, the locations of the map are selecting in decreasing order of importance to analyze the corresponding objects. The normalized bottom-up measure is computed by taking the average difference of the local features with respect to all others map elements, providing a measure of "conspicuity" for each location.

2 The Proposed Model

In this paper an attention system inspired on the human visual attention mechanisms is described, although it doesn't intend simulate it. The aim of this work is to present a new methodology that take the advantages of a bottom-up system in detecting the most different objects in the scenes, while allowing the integration of different features in a set of specific tasks, to be used in the robotic field. The model can be used too to determine the influence among features and objects and the degree of environment interference in the task performance.

2.1 The Feature Maps

The bottom-up, feature-driven, component of attention is provided by the features extracted from the segmented image. Region segmentation is performed with a classical region-growing algorithm that had its bases on some of the Gestalt principles such as similarity and proximity. The method proposed here aims to find in the input scene, objects containing the most "interesting" and "important" information.

The method proposed for detecting such objects is based on the decomposition of the input image into a set of independent features maps. Each map represents the value of a certain attribute computed on a set of objects. Relevant objects can be detected if the corresponding primitives have a feature values strongly different from the other objects present in the scene. Local comparison of feature values are used to compute such measure of "difference" for each feature map and give raise to a corresponding set of *conspicuity* maps.

A set of feature maps is used, where each location (x,y) of the feature map F^k (k=1..K) represents the value of the k-th attribute for the object's center of gravity located at the pixel (x,y). The maps are normalized in the range [0,1] with respect to the maximum value in each feature dimension. In this way, multiple descriptions are available in parallel for each visual primitive and its possible to evaluated a separate measure of interest of each of them.

The attributes used to guide the attention in the system were chosen on the basis of the evidence from the study of human attention[1,7]: average gray level, size, orientation, shape and distance. The feature distance used considers the distance between the object's centroid and the mass center of the scene, and it was chosen to give an advantage to some object in a different configuration from the others (Gestalt) but others criteria can be used.

We based on the theory of Moments[8] to extract the features due its simplicity and invariance properties. By using moments we can extract a lot of useful information such as position, size, orientation and shape. The advantage of moments over other techniques is the implementation of the shape descriptors is straight-forward and they also carry a shape "physical interpretation". The regular moment definition is a projection of the function f(x,y) representing the image in a monomial function $x^p y^q$. The (p+q) bi-dimensional moment for a (NxM) discrete image is defined in equation 1, with (p, q = 0, 1, 2, ...).

$$m_{p,q} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} x^p y^q . f(x,y) \qquad (1)$$

We can normalize the Equation 2 to get invariant moments to translation and scale changes. They are called Central Moments (CMs):

$$\mu_{pq} = \frac{1}{m_{0,0}^{1+(p+q)/2}} \sum_{y=0}^{M-1}\sum_{x=0}^{N-1} x^p y^q . f\left(x - \frac{m_{1,0}}{m_{0,0}}, y - \frac{m_{0,1}}{m_{0,0}}\right) \qquad (2)$$

where $m_{0,0}$ is the object area and the terms $(m_{1,0}/m_{0,0})$ and $(m_{0,1}/m_{0,0})$ are the object's centroid coordinates. The

zeroth order moment $m_{0,0}$ represents the total object area and is given by equation 3.

$$m_{0,0} = \sum_{y=0}^{M-1}\sum_{x=0}^{N-1} f(x,y) \qquad (3)$$

By the second and third order moments, we can get invariant features to the rotation of the objects. These functions, are the Hu Moment Invariants.

The second order moments may be used to determine several useful object features [9] such as the object principal axes, the elongation and the orientation of the major principal axis, given by equation 4.

$$\Phi = \tfrac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right) \qquad (4)$$

where Φ is the angle of the principal axis nearest to the x axis. The orientation of principal axis (θ) specifically may be determined from the values of μ_{11}, (μ_{20} -μ_{02}) andΦ[9]. The orientation (θ) alone does not guarantee a unique orientation since a 180° ambiguity still exists. The third order central moments may be used to resolve this ambiguity.

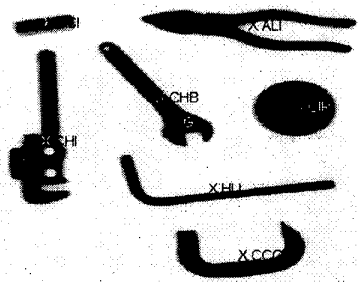

Figure 1 – Scene with objects.

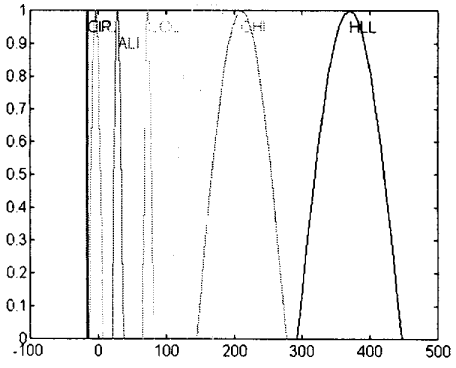

Figure 2 – Gaussian distribution for the sum of the invariant moments for objects in 12 scenes.

As a shape feature, we chose the sum of all Moment Invariant of Hu (S=ϕ1+...+ϕ7) due this sum presents a better values distribution, comparing with any other image set with some isolated moment feature. The figure 1 shows one basic scene with different objects: Integrated Circuit (CCI), Hex Key (HLL), Circle (CIR), Monkey Wrench (CHB), Pipe Wrench (CHI), a pair of pliers (ALI) and a Clamp Iron (CCC).The X foregoing each object is on the center of gravity (centroid) and "G" is the center of gravity of the global scene.

Due to deformation occurred during digitization and the image aspect ration, and due the accomplished transformations (rotating and scaling), the extracted Moment Invariants have a distribution within a minimum and maximum value with a mean and standard deviation showed in figure 2.

2.2 The Conspicuity Maps ($C^k_{,x,y}$)

The measurements represented by the k feature maps F^k are used to detect locations of interest by means of the conspicuity maps. $C^k_{,x,y}$ is related to the difference between the value of $F^k_{x,y}$ and the other values of $F^k_{u,v}$ for the objects located in the others positions and is obtained from the equation 5.

$$C^k_{x,y} = \sum \left| F^k_{x,y} - F^k_{u,v} \right| \qquad (5)$$

The Conspicuity measure $C^k_{,x,y}$ is normalized by the maximum value in each map.

2.3 The Fuzzy Integration Methodology

Fuzzy logic is the process of problem solving that uses imprecise linguistic concepts such as "low", "medium" and "large". The fuzzy logic was proposed by L. Zadeh [10] in 1965, as an extension of the classical sets and it has as a base. The fuzzy set inclusion degrees are determined by functions called "membership functions" (MFs). The main steps involved in the fuzzy process are: fuzzy sets determination, membership functions determination, fuzzification, fuzzy logical operations determination, Inference rules determination, knowledge extraction and defuzzification.

The methodology proposes here is an alternative way to deal to the *binding problem* and to the interference among features combination. A global overview of the system proposed is showed in the figure 3. The fuzzy sets that we used in our work were chosen due their influence in characterizing the bottom-up and top-down attention process and they can be seen in the table 1. The bottom-up salience index (BSI) and the top-down salience index (TSI) are used as input in the integration subsystem that gives a global salience index (GSI) in its output.

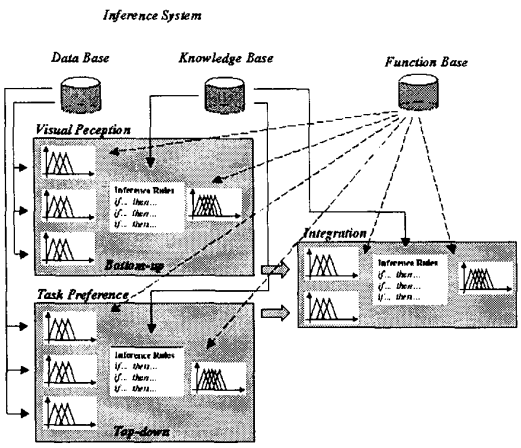

Figure 3 - Overview of the proposed model

2.4 The Bottom-up Fuzzy Subsystem

The bottom-up subsystem was elaborated with 243 rules, combining the five input fuzzy set: "Conspicuity of Gray", "Conspicuity of Size", "Conspicuity of Orientation", "Conspicuity of Shape" and "Conspicuity of Distance" and their respective attributes: "small", "medium" and "large". The output fuzzy set "Bottom-up Salience" has the attributes "small", "small medium", "medium", "medium large" and "large". The indexes BSI, TSI, GSI showed in the table 1 are obtained through a defuzzification method that aggregates the output gotten for the input fuzzy sets through the centroid method [10].

Table 1 - Fuzzy sets of first and second layers network

Fuzzy Subsystems	Input Fuzzy sets	Output Fuzzy Set
Bottom-up	*Conspicuity of Gray*	*Bottom-up Salience*
	Conspicuity of Size	
	Conspicuity of Orientation	
	Conspicuity of Shape	
	Conspicuity of Distance	
Top-down	*Gray average*	*Top-down Salience*
	Size	
	Orientation	
	Shape	
	Distance	
Integration	*Top-down Salience*	*Global Salience*
	Bottom-up Salience	

The shapes of the membership functions (MFs) of the conspicuity fuzzy sets were obtained through the answers gotten from a psychophysics test applied to fifty subjects. The aim of these tests was extract the human knowledge in subjective questions that express qualities of conspicuity of gray, conspicuity of size, conspicuity of orientation and conspicuity of distance.

The results obtained from the tests are used to apply a human perception in a computer vision system but any kind of knowledge can be used. The figure 4 shows a kind of image used in the conspicuity test. In the test the subject was questioned to indicated, with respect to the size of the first object (to the left to the right), which objects are "little similar", "medium similar" or "very similar" to it. In the figure the sizes are normalized within 10 values [1 to 10] or [0 to 1], from the smallest to the biggest object respectively. The conspicuity measures with respect to the object's sizes is normalized in the [0 1] interval through the equation 6 where Cmax and Cmin are the maximum and minimum values in the size scale.

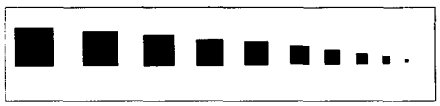

Figure 4 – Image used to determine the membership functions (MFs) of the "Conspicuity of Size" fuzzy set of the bottom-up subsystem.

$$\overline{C}^{k}_{x,y} = \frac{C^{k}_{x,y} - C_{\min}}{C_{\max} - C_{\min}} \qquad (6)$$

In table 2, is showed the results obtained to the "small conspicuity of size". In the figure 5, is showed the results obtained for "small", "medium" and "large" conspicuity of size and the shape of the resulting MFs as function of the conspicuity normalized in the [0 1] interval.

Table 2 – Results for "Small Conspicuity of Size"

Normalized Sizes	Small Conspicuity (% Votes)
1	34 %
2	34%
3	29%
4	3%

Analogous tests were performed to determine the MFs of the other conspicuity sets "Conspicuity of Gray", "Conspicuity of Distance" and "Conspicuity of Orientation" except for the fuzzy set "Conspicuity of Shape" which MFs were adopted. The figure 6 shows the kind of image used in the psychophysics test to determine the conspicuity of gray measures with respect the first object (to the left to the right). The gray levels showed in the gray scale are 15, 47, 79, 95, 159, 191 e 223. The gray levels were normalized within 8 values in the interval [0 1]. The results are showed in the figure 7.

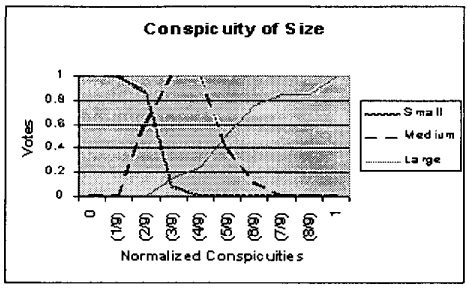

Figure 5 – MFs for the attributes "small", "medium" and "large" of the fuzzy set "Conspicuity of Size" based on the psychophysics test.

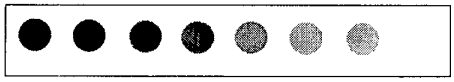

Figure 6 – Image used to determine MFs of the "Conspicuity of Gray" fuzzy set of the bottom-up subsystem.

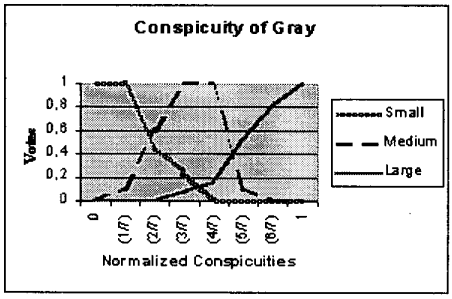

Figure 7 – MFs for the attributes "small", "medium" and "large" of the fuzzy set "Conspicuity of Gray" based on psychophysics test.

The inference fuzzy rules specify the relationship among the fuzzy variables. The rules are specified in t IF-THEN format. These rules are usually specified by a domain expert [10]. Because the bottom-up salience is very influenced by the global context of the objects in the scene (Gestalt influence) some objects can attract attention in a different way if the objects surrounding them change. For example in figure 8(c), if one object has a very different average gray level with respect to the other ones, some other features such as shape and size can not be so relevant in attracting attention. We can perceive also the influence of the distance among the objects in attracting the attention.

In our work, by inferring a set of intuitive bottom-up attentional rules, it is possible to control the output salience through the combination of different conspicuity features present in an specific scene. This flexibility is obtained by fuzzy logic that allows us to weight the conspicuity measures gotten from the (independent) maps, in a context dependent way. By this process, we can get a bottom-up salience index for each object in the scene. This is a partial result that will be processed in the next integration fuzzy subsystem. Figure 9 shows the MFs defined to the output fuzzy set. The attributes are used to describe the system through the inference rules.

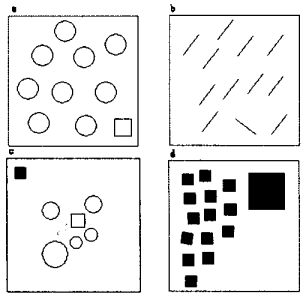

Figure 8 – Example showing how a global scene configuration can be context dependent of the features present in a scene. In (a) the square attracts attention to it because it has a different shape from the others objects and the same occurs in (b) for the line with a different slant. In (c), the difference of gray has a major influence in attracting the attention to the object. In (d) the size has greater influence than orientation. The different orientation square attracts attention to it, but with less intensity.

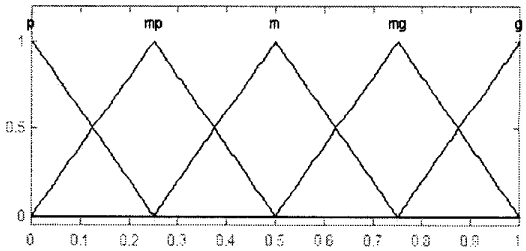

Figure 9 – The MFs for the attributes "small" (p), "small-medium" (mp), "medium" (m), "medium-large" (mg) for the output fuzzy sets "Bottom-up Salience", "Top-down Salience" and "Global Salience".

2.5 The Top-down Fuzzy subsystem

A top-down procedure can be added to guide attention to a specific object in the scene. By definition, top-down attention is related to high-level knowledge depending on task, in terms of what objects must be looked for, or which features should be considered to locate them. In our work these features are based on the object properties such as gray average, size, orientation, shape and distance, but any other features such as color and brightness, slenderness, spread, symmetry, elongation as well as the object relations such as greater (smaller) than, to the left (right) of, above (below) of,

and another one specified by human knowledge. By top-down attention, the computational time in recognizing objects could be greatly reduced.

The second top-down fuzzy subsystem combines different features according to the relevance of them in different tasks, returning us a salience output index for each object present in the scene. The partial result for each specific task can be combined with the bottom-up salience index in the integration fuzzy subsystem.

Again, the fuzzy logic allow us to make decision with great flexibility by using a set of inference rules. By using this strategy, we can specify a set of previous defined tasks stored in a knowledge base. Due the use of five different features, the system is able to perform eighty two different tasks. The top-down fuzzy subsystem is task oriented and it was built with 252 rules for each task, combining the five input fuzzy set that are "Gray"(average gray) , "Size", "Orientation", "Shape" and "Distance" with the output fuzzy set that is "Top-dow Salience". The input fuzzy sets "Gray" and "Size" have the attributes "small", "medium", "large", and the "Orientation" and "Distance" fuzzy input set have the attributes "small" and "large". The fuzzy set "Shape" has seven attributes that represent the shape of the objects showed in the figure 1. The MFs of the "Shape" fuzzy set were defined as function of distribution of values for the shape features (The sum of all Hu invariants) showed in figure 2. The output fuzzy set has the attributes "small", "medium small", "medium", "medium large" and "large". A database was created in order to store different task rules such as: take the smallest a pair of pliers, the darker pipe wrench, the closer and darker clamp iron and so on. In order to build the membership functions of the top-down subsystem, we based in the human perception. The aim of these tests was extract the human knowledge in subjective questions that express perception of qualities of gray (such as dark, bright) size (little, big), orientation (sloping) distance (near, far). The results obtained from the tests are used to apply human perception behavior in a machine vision system but any other kind of knowledge can be used. Figure 10 shows a kind of image used in the perception test. In the test the subjects were instructed to indicated in the gray scale, the qualities "dark", "bright" and "Medium gray". The tests were performed independently from each other in alternate way. The gray scale used shows 256 levels of gray and it is normalized within 10 values [1 to 10] or [0 to 1], from the darker to the brighter level respectively. Figure 11 shows the Membership functions (MFs) obtained through the psychophysics test for the fuzzy set "Gray".

The other perception tests were performed in analogous way, using objects with different normalized sizes, inclination (-90° to 90°) and positions. In the orientation test we just considered the module of the direction but other perception properties, can be included.

Figure 10 – Gray scale used in the perception test.

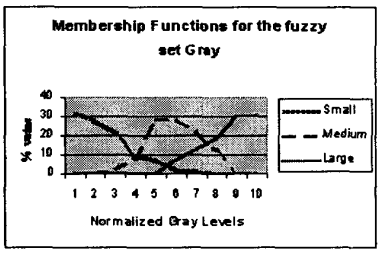

Figure 11 - MFs for the attributes "small"(dark), "medium" and "large"(bright) of the fuzzy set "Gray" based on the human answers in the psychophysics test.

The figure 12 shows the MFs defined for the fuzzy set "Shape" for each object present in the figure 1

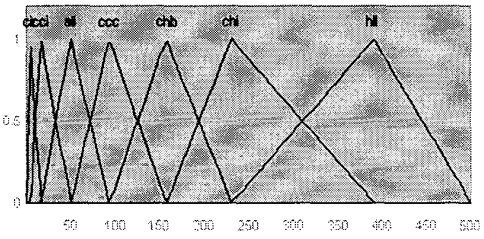

Figure 12 – Key inference based on the distribution of the feature shape values (sum of the Hu's invariants) for tasks involving shapes.

2.6 The Integration Fuzzy subsystem

The aim of the integration fuzzy subsystem is to associate the salience indices gotten from the bottom-up and top-down subsystem giving a global salience result index. The twenty five rules used in this subsystem was specified in order to permit to the bottom-up subsystem enhances or inhibits the result obtained from the top-down subsystem, if an specific task agrees or disagrees with it.

The fuzzy net of two layers simulates a parallelism between the bottom-up and top-down subsystem, however both systems in this work can be considered independent and complementary. The subsystem can be used to guide the top-down process to the most conspicuity object in the scene. The global salience index allow us also to estimate the influence of the bottom-up subsystem with respect to the top-down one

and it can be used to estimate the influence of the context in the task performance and the interference among objects and their features.

3 Experimental Results

In order to verify the system performance in target search tasks, we considered some real scenes such as showed in figure 13. Figure 14 shows the respective conspicuity activity that serves as input for the bottom-up fuzzy subsystem and in the table 3, it can be seen the ISB, IST and ISG indexes.

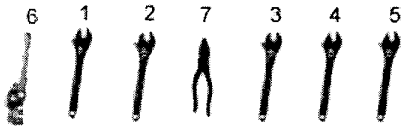

Figure 13 – Example of a scene with objects.

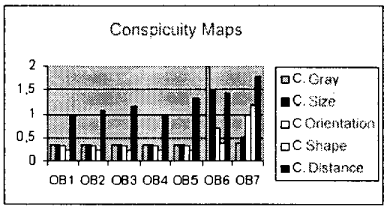

Figure 14 – The conspicuity maps for the objects in fig 13.

Table 3 – Indexes for the task to determine the "pipe wrench slanted and dark"

OBJECT	ISB	IST	ISG
1	0,254	0,448	0,438
2	0,281	0,494	0,49
3	0,333	0,465	0,45
4	0,254	0,469	0,46
5	0,391	0,495	0,388
6	**0,905**	**0,608**	**0,671**
7	**0,754**	0,083	0,19

4 Conclusions

Inspired on the human visual attention mechanisms we developed a new methodology to integrate features within and among the bottom-up and top-down subsystem by a fuzzy net.

The first bottom-up subsystem allow us to combine features and infer with great flexibility some intuitive decision rules based on the visual perception principles such as the Gestalt laws. The second top-down subsystem combines different features according to the relevance of them in different tasks. Finally, the last subsystem aggregates the information of the previous systems and give us a general salience index. The Gestalt theory states that the whole properties are not the sum of its parts, and that each part depends on the context which it is included. Many aspects of Gestalt have been used in Computer Vision but the use of attention mechanisms, a knowledge base in others criteria such as geometrical regularities are rarely considered. The model was tested with several images and the regions of greater attention always corresponded to the target (when it was present). In conflict situations, when the target is not present, and the distractor objects share common features for an specific task, the system was able to take the better decision and select (or not) another object. The division of the net in two layers is important because it help to make decision about the results obtained from the previous layer. Because the global salience index allow us to estimate the influence of the bottom-up subsystem with respect to the top-down one, the new methodology presents here open a further investigation in this field. The determination of the correct weights among features and among rules by using more sophisticated psychophysic tests can be approximate a psychological model for visual human perception and can be used as a metric to evaluate the environment influence in an specific object perception and task performance. The fuzzy logic has some advantages in solving the kind of problem presented in this work, such as its learning facilities, imprecise data tolerance and its proximity with the natural language.

References

[1] Treisman, A. M, Gelade, G., Feature-integration theory of Attention. *Cognit. Psyco*, V.12 (1980) 97-136.

[2] Treisman, A., Features and objects in Visual Processing. *Scientific American*, V.253,1986, 114-125.

[3] Ahmad, S., VISIT: An efficient Computational Model of Human Visual Attention. *PhD Thesis*, University of Illinois, Urbana-Champaign, (1991).

[4] Milanese, R., Detecting Salient Regions in an Image: from biological evidence to computer implementation. *PhD. Dissertation*, University of Geneva, (1993).

[5] Johnston, W., Dark, V.J., Selectivity Attention. *Annual Review of Psychology*, Vol. 37. (1986) 43-75.

[6] Cave, K.R., Wolfe, J.M., Modeling the Role of Parallel Process in Visual Search. *Cognitive Psychology*, Vol.22, (1990) 225-271.

[7] Julesz, B., Towards an axiomatic theory of preattentive vision, In Edelman et al. (eds), *Dynamic Aspects of Neocort. Function*, Wiley, 1984, pp.585-612.

[8] Hu, M.K., Visual Pattern Recognition by moment invariants. *IRE Trans. on Information Theory*, V.8. (1962) 179-187.

[9] Prokop, R.J. & Reeves, A.P., A Survey of Moment-Based Techniques for Unoccluded Object Representation and Recognition. *CVGIP*, V.54, (1992) 438-460.

[10] Zadeh, L., Fuzzy sets. *Information and Control.* Vol.8, (1965) 388-353.