Guiding a Bottom-Up Visual Attention Mechanism to Locate Specific Image Regions Using a Distributed Genetic Optimization

Eanes T. Pereira and Herman M. Gomes

Universidade Federal de Campina Grande, Departamento de Sistemas e Computação
Av. Aprígio Veloso s/n, 58109-970 Campina Grande PB, Brazil
{eanes, hmg}@dsc.ufcg.edu.br

Abstract. The purpose of this paper is to present an approach to locate specific regions in images. The novelty of the approach is the combination of a weighted bottom-up visual attention mechanism with a genetic algorithm optimization running on a computational grid. The visual attention mechanism is based on the model proposed by Itti and Koch [1]. A saliency map indicates the most interesting points in an image using a number of intermediate low level features, which are detected at different scales and orientations. Using the saliency map weights as parameters, the optimization problem is to minimize the number of most salient points needed to locate a set of reference image regions, previously (and manually) labeled as being interesting. Both an objective and subjective evaluation have demonstrated that the proposed approach is more effective when compared to a fixed weight attention mechanism.

1 Introduction

In any physical computational system, processing capability is limited. A mechanism to deal with this drawback in both biological and machine vision systems is visual attention. Visual attention is the ability that the visual system of superior vertebrates have to select and process only the most relevant regions in a visual scene. In this way, only the major areas in a scene are treated. This selection of relevant information in input stimuli is one of the most important characteristics of visual biological systems that allows fast detection of predators and is very important for perpetuation and evolution of the species [1]. Tsotsos [2] analyzed the computational complexity of visual analyses and confirmed that visual attention is one of the most important contributions to optimize the quantity of computations in visual systems.

For study purposes, visual attention can be divided into bottom-up and top-down. Bottom-up visual attention is related to low level features of the scene, such as: color, orientation and intensity. In this case, attention is not a conscious process. Only subjective interest guides the observer attention. Whereas top-down visual attention is very related with the observer desire or purpose. In top-down mechanisms, attention is guided by a previous interest of the observer.

J.F. Martínez-Trinidad et al. (Eds.): CIARP 2006, LNCS 4225, pp. 257–266, 2006.

Generally, models that require a previous process of learning are used to study top-down visual attention. One of the most used are neural networks. To study bottom-up visual attention, one of the most known models is saliency based model proposed by Koch and Ullman [3].

In a saliency map model, a set of maps is combined to form one single map that represents the most salient regions in the scene [4]. A salient region is the region that attracts most attention. There are several ways to combine feature maps. Itti et al [5] compared four strategies to combine feature maps: simple normalized summation, linear combination with learned weights, global non-linear normalization followed by summation and local non-linear competition between salient locations. Almost all strategies used to combine feature maps are based in learning processes to weight the maps. But none of them use an optimization process like genetic algorithms.

As to each saliency map is associated a weight, top-down knowledge can be used to guide the types of selected regions [6]. For instance, if one is searching for red flowers in a garden picture, the search accuracy can be improved if the weight related with color has a higher value than the other features. So, the system proposed here uses that knowledge to improve the quality of the search and guide the attention to look for previously known objects via an optimization process.

This paper proposes a novel strategy for optimizing the weights of a feature based attention mechanism. This strategy uses genetic algorithms to optimize the arrangement of weights that gives the best results when compared with a weightless map. Optimization is followed by detection and comparison phases. In the detection phase, a saliency map based on three features (color, intensity, and orientation) is constructed. After the detection phase, the resulting salient regions are compared with some previously selected regions and the comparison results are used in the next optimization phase.

The problem of mapping an image region that raises the human visual attention into a function that could be optimized is a complex task. Even when using the linearly weighted saliency model used in this work, there is not a clear way of doing that mapping. Besides that, the quantity of possible feature map combinations is very large and adds more complexity to the problem treated here. Therefore, we chose to use a method that has been used to treat those types of problems: genetic algorithm optimization, which are most appropriate for optimizing complex models in which the location of a global optimum is a difficult task [7].

There are many works that use some kind of evolutionary or genetic approach to treat visual attention [8], [9], but none of them uses a genetic algorithm to weight features maps like the strategy proposed here. Stentiford [8] presents a strategy that maps pixel neighborhoods to individuals in a genetic population. This population is evolved and performs a discrimination between salient and non-salient image features. Treptow et al. [9] present an evolutionary algorithm that uses the Adaboost framework to find new features and to reduce feature search.

The paper starts with a description of the overall optimization architecture proposed, which includes a description of the visual attention mechanism that was employed. This is followed by an objective (numerical) and subjective (based on some test images) evaluations. Finally, some conclusions and proposals of future work are presented. This paper presents and extension of a previous work [10] by giving an improved description of the proposed approach and providing an expanded set of results that uses images publicly available.

2 Proposed Approach

The approach receives as input a reference set of static color images, digitized at 352 × 240 pixels. The images are grouped in two sets: one containing only objects and other with people. In the set that contains people we consider the faces like the regions that raise more attention. So, in these images, human faces are marked manually by the selection of ears, eyes, mouth and nose of the present faces in the image. After that a file containing the coordinates of these parts is created. In the images without people, all the objects or regions that would intuitively raise the attention of human observers are also manually identified. These images are transfered to remote machines in a grid during the optimization process.

Since the size of the input space for optimization is relatively high (a total of 27 continuously valued weights), a computational grid was employed to cut down the processing time from several weeks on a single computer to just a few days using a grid of dozen computers. The reference input images (100 for people and 80 for objects) are divided into subsets for grid processing. The detection module (which is actually the attention mechanism) will concurrently run on different machines of the grid and process each of the reference image subsets producing as output a list of the most salient points found on those images. A genetic algorithm (which is the optimization module itself) provides the different weight combinations for the attention mechanism. The algorithm uses a cost function that is the percentage of salient points needed to find the previously labeled regions on the reference images. This function is minimized throughout the several iterations of the algorithm. Figure 1 illustrates the whole process.

2.1 Optimization Module

The optimization module is composed by a genetic algorithm that generates sets of weights to be applied in a combination of feature maps. An initial population of weights is randomly generated. After that, this set is sent to remote machines. A group of images and the detection module is sent to remote machines too. In the remote machines, the detection and evaluation processes take place.

When the detection and evaluation processes are finished, the results are sent to the optimization module. Then, the genetic algorithm tries to minimize the number of points needed to find a set of previously selected regions.

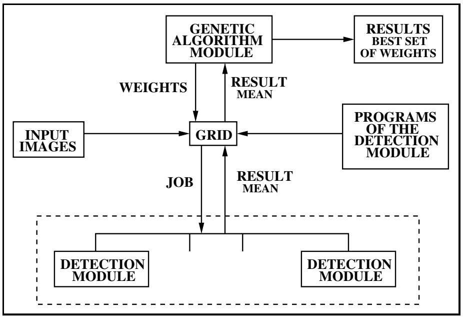

Fig. 1. Optimization architecture of the proposed approach

2.2 Detection Module

The detection module is an adaptation of the model proposed by Itti et al [11]. It uses a saliency-based attention mechanism (bottom-up), which is constructed from a Gaussian Pyramid and locally oriented neighborhood operators. Figure 2 shows a diagram of the detection module.

Initially, three types of primitive visual features are extracted: color, intensity and orientation. After that, four color channels are created (R to red, G to green, B to blue and Y to yellow). Finally, for each channel a Gaussian Pyramid with five levels is created. The Gaussian Pyramid is composed by pass-low filtered versions of the Gaussian convolution applied to the input image. The pyramidal representation is used to get image samples that do not have undesirable details.

To obtain the center-surround differences, it is necessary to create Steerable Pyramids. A Steerable Pyramid is a multi-scale and multi-orientation decomposition of an image. In this type of decomposition, an image is subdivided in a set of sub-bands localized in different scales and orientations. Center-surround operations are implemented as differences among scales. The center is a pixel in the scale $c = \{2, 3, 4\}$ and the surround region is the correspondent pixel in the scale $s = c + \delta$ with $\delta = \{3, 4\}$. Difference between two images is obtained by image interpolation in the scale and point-to-point subtraction. The

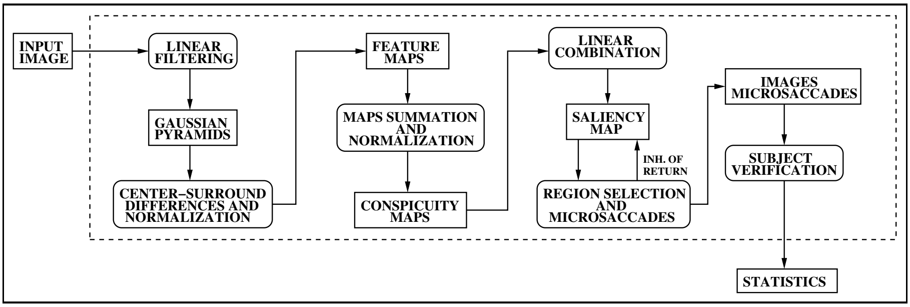

Fig. 2. Detection module

utilization of several scales allows multi-scale feature extraction. After execution of center-surround differences, the feature maps are generated.

Once the feature maps are obtained, they are summed up to produce the Conspicuity Maps: $\overline{\mathcal{I}}$ for intensity, $\overline{\mathcal{C}}$ for color and $\overline{\mathcal{O}}$ for orientation, in the scale $\sigma = 4$. The motivation for create three separated channels ($\overline{\mathcal{I}}, \overline{\mathcal{C}}, \overline{\mathcal{O}}$) is the hypothesis that similar features compete for saliency, while different features independently contribute for Saliency Map [1]. The three Conspicuity Maps contribute to saliency. The purpose of the Saliency Map is to represent salient regions in the image by scalar quantities and guide selection of regions based on spatial distribution of saliency.

Before summing the Conspicuity Maps, the set of weights got from the genetic algorithm are applied. Each combination is applied to all images and the results are written to a file. These weights are applied to the Conspicuity Maps, Feature Maps and Saliency Maps. After the subject detection, the results are compared with the previous manually selected ones. The result of this comparison is returned back to the optimization module.

A pixel sorting scheme (in descending order of saliency) was implemented to select regions of interest. A region around a coordinate of interest (which corresponds to the pixel with greater value) in the Saliency Map is selected. The radius of this region is called by the inhibition radius. In the experiments presented in this work we used inhibition radius of 2 pixels. Besides selecting the region of interest, that region is filled with null intensities. This prevents that the same region of interest be treated another time.

To prevent the same region of interest being selected more than one time, only part of objects are located, a micro-saccadic movement strategy was implemented. For each region of interest displacements are executed, changing the focus of attention to several neighboring points.

A previous work [12] employed an averaging computation in order to obtain the final Saliency Map, and to perform the required visual task using this map (e.g. locating traffic signs, locating faces, etc). In this paper, however, we tune the attention system by changing a set of weights that are used to produce the final attention map, in such a way that the given visual task is performed better. These weights are obtained by an experimental process. In that process, different weights are attributed to each map and results are optimized by the genetic algorithm until it stops.

2.3 Labelling the Regions of Interest

In order to guide the optimization process and to verify if the results of the automatic detection module are satisfactory, one tool was implemented to help manually select objects in images. Satisfactory results are those in which a smaller number of most salient points are required to fall into a set of previously selected image regions.

2.4 Verification Process

After the optimization module achieved a convergence plateau and returned the set of optimized weights, a verification process is performed. That process is performed with images obtained from Internet websites and that were not used in the optimization process.

The verification process is described as follows. Regions of images that raise attention of the observer are manually selected and their coordinates are saved. After that, the visual attention system is applied to those images and the coordinates of all points that raise attention are saved. Then a program is used to verify if the points returned by the attention system are contained in the regions manually selected. This process is done in two ways: using the optimized weights and without using them.

3 Experiments and Results

The purpose of the experiments was to optimize the weights in such a way that the system could find the subject of interest in the images with the least number of points. As mentioned before, the experiments were done using two types of images: with and without people. The goal of using different types of images is to check the generalization capacity of the genetic algorithm.

In this work, grid computing and a genetic algorithm library were used. The computational grid framework *OurGrid (http://ourgrid.org/)* was used to remotely process the detection module. The *GAlib (http://lancet.mit.edu/ga/)* genetic algorithm library was used to make our system. Almost all the code was done in C++ language, the only exception was the manual detection module and the methods related to the grid communication that were done in Java language.

The optimization (or reference) set of images containing people has 100 images and the set of the images without people has 80 images. These images were obtained in an environment that contains great quantity of dispersive elements that can misguide the subject detection process. The image set without people is formed by indoor images and natural environments.

The genetic algorithm uses overlapping populations. Using a previous established percentage, the algorithm creates a new population from a percentage of the best individuals in the early population and from a percentage of the crossovers and mutations of the early population. The algorithm goal is to determine the best mean of points needed to find all previous manually selected regions.

After the optimized weights were found, a verification process was performed. This process was done using 100 images with people and 100 images without people that were not in the optimization set. These images were obtained from Internet websites. In the image set with people, faces regions of people were manually selected. In the other image set, regions that raise attention in accordance with some features (color, intensity, and orientation) were selected. After that, a verification of the most salient points found by the visual attention system that was contained in the faces regions was performed. Fig. 3 shows the results obtained using an inhibition radius of 2 pixels.

The image set with people used in the verification process contains 194 people faces. From the graph of Fig. 3 (left), one can see that using only 1% of the total

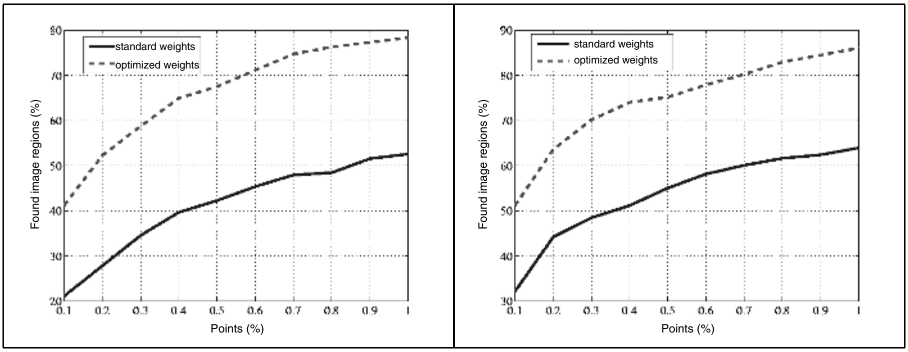

Fig. 3. Region location results for images with (left) and without (right) people, when using the corresponding optimized saliency map weights

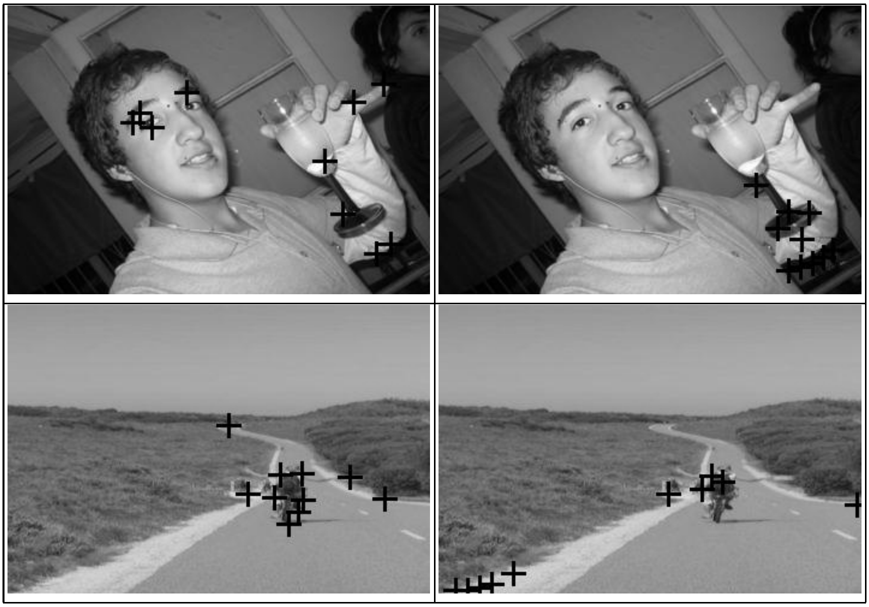

Fig. 4. Results of the attention mechanism with optimized (1st column) and standard (2nd column) weights on arbitrary test images

number of image points the system using optimized weights found points of interest in 152 (78%) people faces previously selected. In the image set without people, 258 objects, or regions that raise attention, were manually selected. The graph from Fig. 3 (right) shows that using 1% of the total number of image points the system using the optimized weights found points of attention in 222 (86%) objects or regions.

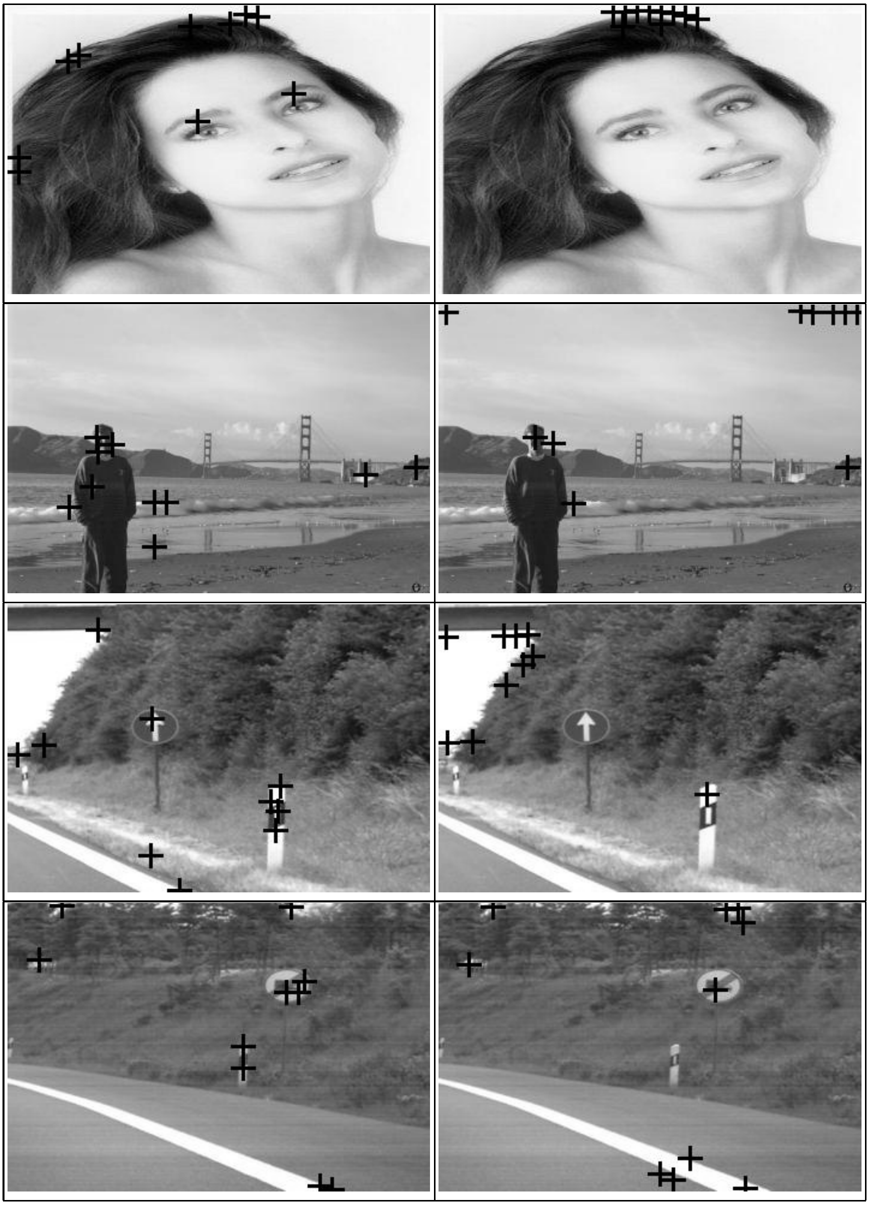

Fig. 5. Results of the attention mechanism with optimized (1st column) and standard (2nd column) weights on test images available at http://ilab.usc.edu/imgdbs/

In Fig. 3, it is clear that the use of optimized weights gives an improvement in the task of finding the subject of images. Besides that, optimized weights guide the subject detection in such a way that the user can previously establish what kind of objects he wants. The graph from Fig. 3 (left) shows an improvement of about 20% to the minor number of points and of about 30% to the greater number of points, and the graph from Fig. 3 (right) shows an improvement of about 19% to the minor number of points and of about 23% to the greater number of points.

For a subjective evaluation, we have applied our optimized attention mechanisms to a number of test images. Figures 4 and 5 contains these results. In Fig. 4, the images were arbitrarily chosen from the Internet. In Fig. 5, we used images from the *Miscellaneous artwork, posters and portraits* (first image) [11], *Miscellaneous outdoors* (second) and *Color images with German traffic signs* (third and fourth images) [13], all available for download at *http://ilab.usc.edu/imgdbs/*. The black crosses in the images indicate the 10 most salient points returned by the visual attention system, using an inhibition radius of 10 pixels. The inhibition radius of 10 was used to help the subjective analysis of the images (reducing the excessive agglomeration of salient points). In these figures, there are two types of images: Images in which the regions of interest are the faces of people, and images in which the regions of interest are general objects. Comparing the results with and without optimized weights one can perceive the improvement brought by the former.

4 Conclusions

This paper presented the application of genetic algorithms in a visual attention system. The genetic algorithm was used to optimize weights that were applied in a saliency map system. These weights were applied to construct the Saliency Map. To obtain results in satisfactory time, the experiments were executed in a computational grid due to the number of free parameters to optimize.

Although the saliency map system used was simple (using only 3 features), the results were very satisfactory. The results using the optimized weights presented an improvement of about 20% against the results of the system without optimized weights. Another characteristic of the experiments presented is that all images used were obtained in arbitrary (mostly distracting, full of low level details) environments. These environments acted misguiding the system in some cases.

Experiments and results shown here have interesting areas for further improvement: only three features were considered in the saliency map model (more features could be added in the future), and there are many aspects of the evolutionary process that could enhance these results, including, for instance, selection methods, diversity maintenance, multi-objective techniques, and the use of automatically defined functions. Future work will further examine these issues and their application to visual attention systems.

Acknowledgements

We would like to thank HP-Brazil and CNPQ (Brazilian Research Council) for their support.

References

1. Itti, L., Koch. C., Computational Modeling of Visual Attention, *Nature Reviews Neuroscience*, 2(3), 2001, 194-203.
2. Tsotsos, J. Analyzing Vision at the Complexity Level, *The Behavioral and Brain Sciences*, 13(3), 1990, 423–445.
3. Koch, C., Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry, *Human Neurobiology*, 4, 1985, 219-227.
4. Itti, L., Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention, *Vision Research*, 40(10-12), 2000, 1489-1506.
5. Itti, L., Koch, C. A comparison of feature combination strategies for saliency-based visual attention systems, *Proc. SPIE human vision and electronic imaging IV*, San Jose, USA, 1999, 473-482.
6. Itti, L. Models of Bottom-Up Attention and Saliency, In: Neurobiology of Attention, (L. Itti, G. Rees, J. K. Tsotsos Ed.), Jan 2005, 576-582, San Diego, CA:Elsevier.
7. Mardle, S., Pascoe, S, An overview of genetic algorithms for the solution of optimisation problems, *Computers in High Education Economics Review*, 3(1), 1999.
8. Stentiford, F. An evolutionary programming approach to the simulation of visual attention, *Proc. Congress on Evolutionary Computation*, Seoul, Korea, 2001, 851–858.
9. Treptow, A., Zell, A. Combining Adaboost learning and evolutionary search to select features for real-time object detection, *Proc. IEEE Congress on Evolutionary Computation*, Portland, USA, 2004, 2107–2113.
10. Pereira, E., Gomes, H., Florentino, V. Bottom-up visual attention guided by genetic algorithm optimization, *Accepted to IASTED International Conference on Signal and Image Processing*, Honolulu, USA, Aug, 2006.
11. 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.
12. Siagian, C., Ititi, L. Biologically-Inspired Face Detection: Non-Brute-Force-Search Approach, *In: First IEEE-CVPR International Workshop on Face Processing in Video*, Jun 2004, 62-69.
13. Itti, L., Koch, C. Feature Combination Strategies for Saliency-Based Visual Attention Systems, *Journal of Electronic Imaging*, 10(1), Jan 2001, 161-169.