Moving Object Segmentation using Visual Attention

Emerson J. Olaya and L. Abril Torres-Méndez
Robotics and Advanced Manufacturing Group, Cinvestav Saltillo
Ramos Arizpe, Coah, 25900, Mexico
abril.torres@cinvestav.edu.mx

Abstract

We have developed a visual attention algorithm that combines previous existing methods to solve the problem of segmenting moving objects in real time. Our approach allows us to select regions of interest over which we force the fixations on a vision mechanism. We create saliency maps with characteristics that highlight within the scene. The amount of maps that can be extracted in an image is huge, so we just use some to avoid high latencies that can harm the performance of our system. Our approach is of special interest when there is no specific object to look for by the system. Thus, a scene is explored in a more natural way compared to simply sweeping out in some order the scene point by point or to wait patiently that an object of interest appears. With this, we assure that all relevant visual information in the scene is taken into account according to the priorities and objectives of the system.

1 Introduction

Segmentation is one of the fundamental problems in image processing. For humans and most animals this task is relatively straightforward. For machine vision systems, however, segmentation can be a very complex task due to the mechanisms involved and the fact that different interpretations about what to segment may exist. Most image retrieval algorithms are based on the global features extracted from static images [14]. However, when a user is interested in retrieving a moving object from a set of images (i.e. video frames), the need of extracting reliable local features, i.e., segmenting the object from its background, turns to be a difficult task. For the particular problem of segmenting moving objects in real time, also known as active segmentation [10], it is relevant to pay attention to the objects of interest in the scene and choose a fixation point within the area or region that covers them. Once a fixation point is chosen, the surrounding visual characteristics can be easily extracted and grouped according to its properties. Existing segmentation approaches [9, 2] assume that a fixation point from where to start the segmentation is given. However, using a visual attention model can solve the problem of selecting automatically a fixation point. The fixation points selected should be ideally located on the object of interest and close to its center to facilitate segmentation.

Distinct science areas, such as psychology, physics, cognitive neurology and computer science, have studied visual attention for more than a century. Our approach is based on the field of cognitive neurology, where one of the main goals is to study the eyes movement to obtain information about the human visual attention. However, as we are dealing with artificial vision systems, rather than focusing on the eyes (cameras), we need to focus first on deciding which visual characteristics may be more relevant within a dynamic scene and then direct the cameras to them. We explore the entire scene in a more natural way compared to simply sweeping it out point by point in some order or to wait patiently that an object of interest appears. With this we assure that all relevant visual information in the scene is taken into account according to the priorities and objectives of the system.

Luis Enrique Sucar and Hugo Jair Escalante (eds.): AIAR2010: Proceedings of the 1st Automatic Image Annotation and Retrieval Workshop 2010. , volume 1, issue: 1, pp. 35-46

The outline of this paper is as follows. In Section 2, we describe related work on visual attention models. Section 3 briefly describes the artificial visual system we constructed. In Section 4 we give details of our visual attention approach. Section 5 describes the saliency map generated together with the experimental results. Finally, in Section 6 are the conclusions and future work.

2 Related Work

There exist different methods in the literature for visual attention. We propose a visual attention approach based on a combination of three existing methods. First, there is the approach used by von Helmholtz [17], who observed the relation between the eyes and an involuntary segmentation process of the visual field, where the eyes are attracted to objects that have not already been tracked. In other words, we involuntarily select regions of the space (the "where") based on the visual characteristics which are, generally, outside the fovea, and then, under the approach of James [6], we fix voluntarily our attention over the selected region (the "what"), with the goal of identifying it, exploring it or just not losing it from the sight. These two approaches were reinforced by Nakayama and Mackeben [11], who gave evidence of this dichotomy in the attention, first the quick and transitory aspect and then the slow and steady one. In the 80's, Klein [7] presented evidence of a new component in the attention called inhibition of return which consists in a type of selective attenuation of regions on a saliency map, avoiding that the focus of attention is directed to regions already visited. This new component has certain similarity with the Helmoltz's approach from the point of view that when eyes are attracted to unknown or new regions is somewhat equivalent to be repelled from already explored regions.

We have mentioned about fixing the sight in a point of interest in the scene, but the next question arises: How to choose the point of interest within an unknown scene? Humans and some animals solve this problem by using a very powerful biological tool known as visual attention. As established by Itti [5], the main purpose of visual attention is to direct the sight to objects of interest; it is for this reason that visual attention and eye movements are closely related. Therefore, the ability to visually understand or interpret a scene goes together with the object recognition problem, which restricts the selection of the regions that must be attended. Based on this, it is established that people use the combination of two approaches: the bottom-up approach, in which the direction of sight is determined by using relevant visual characteristics based directly on the visual information; and the top-down approach, where visual cues are used depending on the task to carry on (e.g., exploring, tracking or searching objects of interest, etc.) The bottom-up approach is based on the hypothesis that certain visual characteristics (i.e., pre-attentive ones) inherently attract the attention (e.g., color, contrast, intensity, edges, etc.) The top-down approach requires additional information to establish the preferences in the estimation of the visual attention map.

In addition to the characteristics mentioned above, there exist other similarities between our artificial visual system and the human visual system. One is the geometric configuration of the cameras, i.e., we have the coplanarity restriction (via software) between the optical axis of the cameras and a point of interest within the scene. Other characteristic of particular importance is focused on the motor abilities of the human visual system [16]. Our system was designed with kinematic abilities similar to the human eye, but with different dynamics.

3 The visual system design

The first aspect to consider when designing an active vision system is the type of configuration we want our system to have. We have implemented an active visual system based on the Fick architecture [4] similar to that presented in [13] Through a visual servo structure we can force the fixation point on static

or moving objects by using a vision system with Fick architecture [4] of 4 DoF for camera motion. After the fixation point is reached we estimate its 3D location using extra-retinal signals [18], which come from the rotational encoders that are used to calibrate and modify the binocular disparity [1]. In Figure 1 we illustrate the final physical assembly of our active visual system. We highlight that the similarity of our artificial visual system with that of a human, far from being anthropomorphic, is more in terms of its qualities of being active, its geometry and functionality.

Figure 1: The active visual system constructed for this research work.

4 The visual attention algorithm

We have considered the distinct methods and ideas mentioned in Section 2 to develop our visual attention algorithm. The algorithm allows us to select regions of interest over which we forced the fixations of a vision mechanism. We then create saliency maps with characteristics that are highlighted within the scene. The amount of maps that can be extracted in an image is huge, so we just use some of them to avoid high latencies that can harm the performance of our system.

In order to select an object of interest within the image scene we have developed a typical bottom-up visual attention model based on the Itti's model [5] (see Figure 2), with the only difference that we use foveated images. We associate a degree of preference or weight to each of the extracted characteristics (saliency maps) such as movement, color, distance to the center of the fovea, etc. These weights can be modulated as a function of the task to be carried. For example, if the task consists in following a red object, we give more weight to visual cues of color and movement and inhibit the characteristic of contrast, depth, illumination, etc. Once a point of interest is selected over any of the two images obtained from our active stereo system, we need to solve the correspondence problem. With this information we can calculate the desired position to force the fixation point over the point of interest. A Kalman filter [19] is used for the case of tracking, this filter predicts the future position of the tracking object based on previous observations and the model of the object is used to recognize the object in the other image.

4.1 Visual abilities

Our active stereo system is capable of fovealizing points of interest or previously known objects within the scene by using a conventional CCD and an exponential retinotopic mapping (explained in Section 4.2) to resample the images. Working with foveated images improves significantly the time of extracting information from the images. However, its variable resolution makes eye movements necessary to drag the projection of the object to the fovea at each retina (camera). One of the key characteristics of active vision is its real-time requirement. Having cameras able to move allows searching for new strategies to decrease the response time. This is the main motivation of working with foveated images. One of the objectives of a system with foveated vision is to achieve a good combination between a big aperture

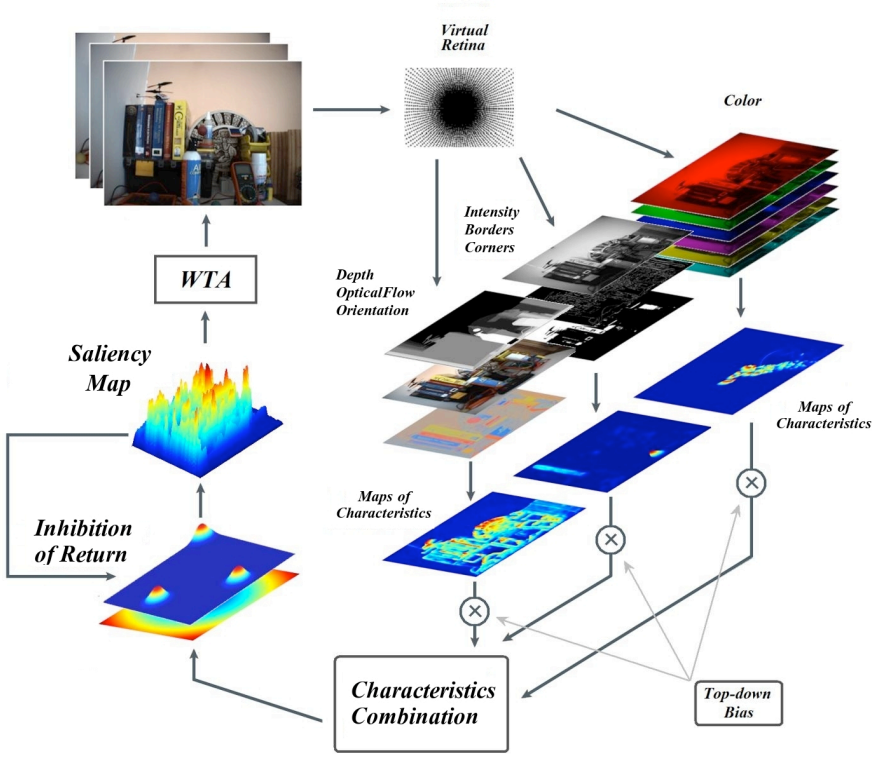


Figure 2: A diagram of the proposed bottom-up visual attention model that uses foveated images. Adapted to our model from Itti's model [5]. First, the virtual retina algorithm is applied to the input images. Second, from the foveated images we extract some characteristics (color, motion, etc.), obtain their map and associate a weight to each of them depending on the task to be carried. After that a characteristics combination step is performed followed by a filtering step that inhibits those already visited places on the image. Finally, a saliency map is constructed which is used to segment the moving object of interest according to the Winner-take-all (WTA) algorithm.

angle of the camera with a significant decrease in the number of pixels, reaching the maximum resolution over the regions of interest (fovea) [12].

4.2 Retinotopic mapping

At a biological level, the term retinotopic means that near points to the scene that are projected into the retina are mapped near the striated cortex, i.e., the retinal topography is respected. Although it is presumed that in primates the retinotopic mapping is polar-logarithmic [15], we have defined our retinotopic mapping as an inverse Cartesian mapping. It is an inverse mapping because we make it exponential from the memory to the image space using the following equation:

$$X = x + \mathrm{sgn}(x)\lfloor K^{|x|} \rfloor - 1, \tag{1}$$

where $\lfloor w \rfloor$ is the floor function of $w, w \in R$; X is a vector that represents the coordinates of a pixel (U,V) in the original image, x corresponds to the coordinates of the pixel in the compressed image (u,v), where $\{X, x \in Z^2\}$; and $K > 1 \in R$ is a constant. In this way, a pixel (u,v) given in the compressed image takes

the corresponding value of the pixel (U,V) obtained from Equation 1, as follows:

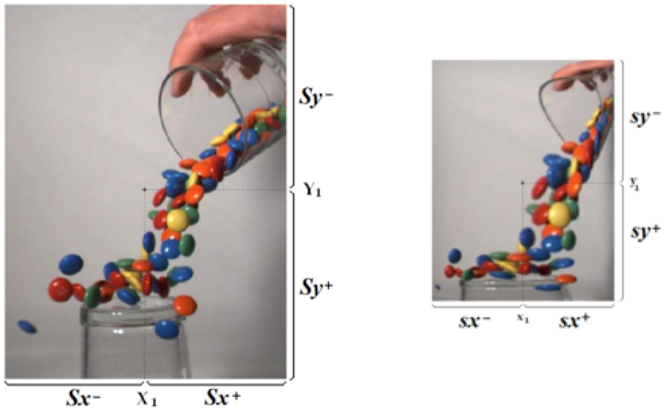
$$[U,V] = [u + \lfloor K_1^{|u|} \rfloor - 1, v + \lfloor K_2^{|v|} \rfloor - 1]. \quad (2)$$

Equation 1 describes an infinite family of functions to resample the image. The exponential component establishes how the resolution diminishes according to the pixel distance to the fovea. Therefore, we need to define K carefully so the image is resampled including the borders and the aperture angle of the camera is preserved without going beyond the limits of the image.

A good way to define constant K in a general form is:

$$K_x = (Sx/2 - sx/2 + 1)^{sx/2}, K_y = (Sy/2 - sy/2 + 1)^{sy/2}. \quad (3)$$

Where Sn and sn (see Figure 3) represent the size of the positive and negative axis of the original and compressed images when the origin is displaced to pixel (X_1, Y_1). Equation 3 is used to design a mapping that decreases the area of the images coming from the camera to a quarter of the total size, i.e., from 480×640 to 240×320, with the origin (fovea) in the center of the image. In this way, we obtain $K_x = 161^{(1/160)}$ and $K_y = 121^{(1/120)}$. The complexity of this algorithm can be inferred from Equation 1, which selects the pixels from the buffer with which a new image is formed to be processed. Although this image is smaller, it preserves the same aperture angle of the original image. This allows us to have an effect as shown Figure 3b, by reducing the image while keeping the retinotopic property of the sensor.



(a) (b)

Figure 3: (a) Original image, (b) the compressed image. Note in (b) how objects near the fovea keep their original size while far objects decrease in size and resolution.

4.2.1 Segmentation using visual attention

The ability of the human brain to process images is far superior to any computational algorithm created. For example, it is excellent extracting the color of an object even with the presence of external variables, such as illumination. When we segment objects many different cues or characteristics are extracted. Color is one of the most relevant and commonly used to segment objects. The color of an object not only depends on the chemical composition of its surface, but also on the conditions of environment: illumination, intensity, number and color of the illumination sources, location and shape of the object, the intrinsic and extrinsic of the sensors, etc. Based on this knowledge we have developed a simple but fast algorithm that keeps the color constancy of objects. The procedure consists on changing the RGB format of each color pixel (of the fovealized image) to its quaternion representation assuming linearity in the reception of the luminous spectrum of the accopled device of the camera charge upon

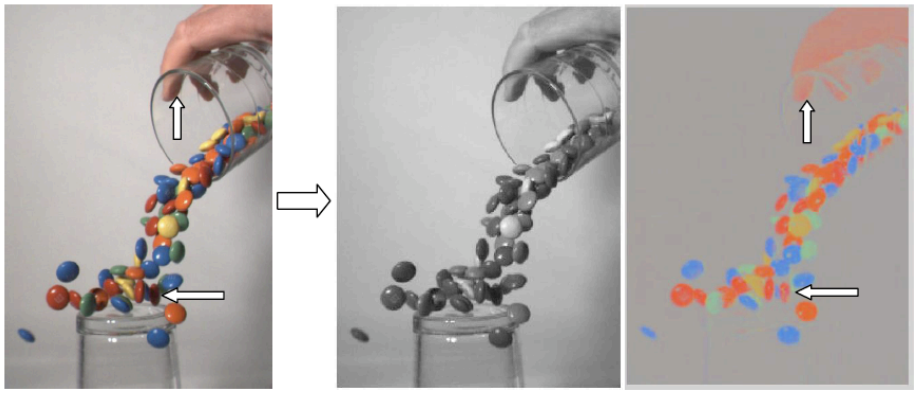


Figure 4: Visualization of an image in its quaternion representation. Left image is the original image and to its right is the quaternion representation: the grayscale intensity w and the vectors of color directions. Note how the colors look more homogeneous, eliminating the shading produced by the illumination conditions and objects geometry.

illumination changes. This representation helps in the segmentation because the direction of the color vector $x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$ is less sensitive to illumination changes, thus facilitating the segmentation using color thresholding. The quaternion representation is defined as:

$$w = x\mathbf{i}+y\mathbf{j}+z\mathbf{k}, \tag{4}$$

where w is the intensity of the pixel, and can be defined with the color components on the RGB triangle:

$$w = \frac{\sqrt{R^2+G^2+B^2}}{\sqrt{3}},\ x = \frac{R}{R+G+B},\ y = \frac{G}{R+G+B},\ z = \frac{B}{R+G+B}, \tag{5}$$

where R,G,B, correspond to the red, green and blue components, respectively. Therefore, we say that a pixel $P_{u,v} = w_{u,v}+x_{u,v}\mathbf{i}+y_{u,v}\mathbf{j}+z_{u,v}\mathbf{k}$, where u and v are the pixel coordinates within the image, will be of a color of interest if its vector of directions $v = (x,y,z)$ is additive inverse of the color being searched $P = 0+x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$, i.e.:

$$(P_{h,k}-P) \simeq w_{u,v}+0\mathbf{i}+0\mathbf{j}+0\mathbf{k}. \tag{6}$$

The result of converting an image to its representation in quaternions is a format not well understood by the computer as an image, but it can be seen if we decompose it in two: the magnitud w and the direction $(\mathbf{x}i+\mathbf{y}j+\mathbf{z}k)$ as it can be observed in Figure 4. It can be noted that the quaternion representation makes the colors less sensible to variations in illumination. Note that same color objects but with different illumination in the left image are seen practically identical in the right image. Although this does not solve completely the problem, it will facilitate in great measure the search of the object of interest. In other words, what we do is to classify colors by grouping colinear vectors in just one vector.

4.2.2 Obtaining the color histogram

By using Equation 5, which implies that $(x+y+z) = 1$, we can project the color on one plane (RGB triangle) and form histograms like the one shown in Figure 5. To obtain these histograms, the first step is to locate each color in the plane, there exists infinite ways to do this. In Figure 5, it can be observed that the red color is located in coordinates $(255,0)$, the blue in $(0,255)$ and the green color in $(383,383)$. The easiest way is to put the blue in the origin, the red in $(0,255)$ and the green color in $(255,128)$. This way a pixel's RGB components are located in the coordinate plane $(x,y) = (g,b+g/2)$. The following step is to define the desired resolution in the RGB triangle. The parameter we choose to restrict the resolution

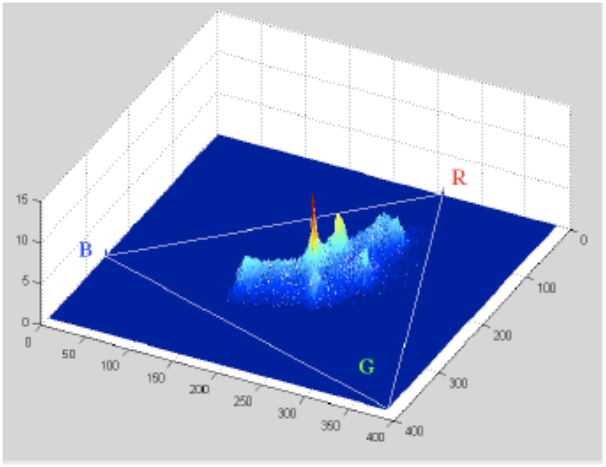
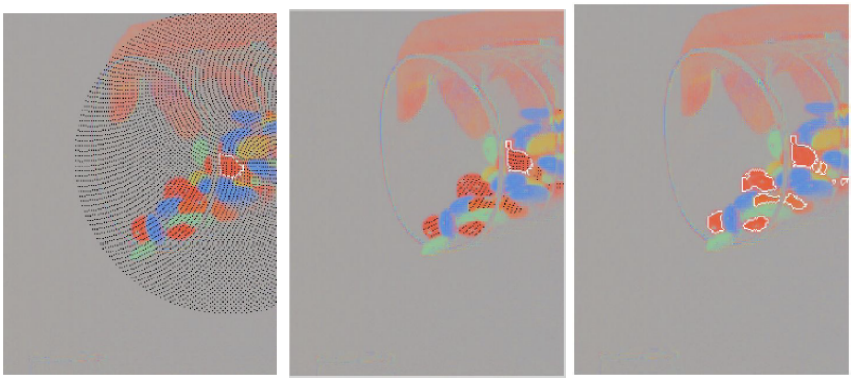


Figure 5: Histogram represented on the RGB triangle of the image in Fig. 4. The red color is located in coordinates $(255,0)$, the blue in $(0,255)$ and the green in $(383,383)$. By using Eq. 5, each color pixel on the image can be projected on the one plane (i.e. the RGB triangle).

is the number of transitions (n) between two primary colors. The last step is to count how many pixels fall in each of these slots and make a graph.

4.2.3 Searching for the object of interest

The search of the object of interest is based on the local extracted characteristics by the visual attention algorithm. The algorithm first selects the point of interest within the scene and then extracts some characteristics of that area, such as its histogram, the contrast map and the rate of change (explained in the next Section). We describe here the algorithm used to extract geometric characteristics of an object. As an start we assume that the object has a characteristic or set of characteristics C, as color, intensity, texture, etc. The algorithm begins by sweeping the images from top to bottom in search of C. At the moment in which this characteristic is found in a pixel $P_{u,v}$, it starts surrounding the pixel by searching all 8-connected neighboring pixels in clockwise direction in order to obtain its silhouette.



(a) (b) (c)

Figure 6: Experimental results of the polar searching of objects. (a) shows the area of the pixels to search; (b) presents in black the pixels that has characteristic C; and (c) shows the objects found "of relevant size" for which their silhouette was extracted.

In this way, we can obtain the contour of a set of pixels with a common characteristic in a fast way and without using derivative-based operators that do not guarantee a closed curve and require slim filters that are sensible to the borders that the object may have inside. Once the silouhoutte of the object is obtained in a vector form, we can extract useful information to achieve the recognition of an object model previously saved in memory, such as the area (moment of order zero), perimeter, shape (compact or regularity factor), centroid, etc. If we look for an object that has been found in previous frames, it is not necessary to search the whole image. If we know the sampling rate of the camera and the maximum velocity of the motors, we can generate a search radius from the last position where the object was seen. In our case, we make the search in a spiral form increasing the radius r by two pixels each 2Π radians. The angular increment is given by $\frac{\pi}{r-1} + 0.01$.

Continuing with the example of Figure 4, we show now experimental results of the search algorithm described above in Figure 6. In this case the characteristic C to search is the *red color*. It can be observed that the algorithm is robust enough to detect in a predefined search zone the objects that contain the characteristic to search.

5 Saliency map

On of the most successful models in computational visual attention was proposed by Koch and Ullman [8], who based their model in a certain type of topographical map, totally compatible with previous presented approaches. They associated a saliency measure, based on the extraction of visual cues to each region of the image forming with all of them a bi-dimensional attention map known as *saliency map*. To build a saliency map we need to integrate the information extracted from multiple visual cues, such as: color, geometry, optical flow, intensity, etc., known as pre-attentive [5], with which we calculate the maps with relevant characteristics, as the contrast in color and intensity, motion, color histograms, geometry, etc. Then, all these characteristics are weighted depending on the task to solve and combined to form the saliency map, then the winner-take-all algorithm is used to select the most relevant region in the map. In the following sections we describe the algorithm to obtain the saliency map.

5.1 Contrast and color maps

As users we can define a priority or degree of interest for the different colors based on the task. We also want our robot to have a certain degree of preference for some color or set of colors. Therefore, we set a weight different to zero to this map only when there is a search task.

Contrast is defined as the difference in color and intensity between a region and its surroundings. We can literally use this definition to generate the contrast map, which can be thought as a 3D filter, either a high-pass filter or magnitude of the gradient. The contrast map, Mc, is the result of the convolution in the space domain for the image $I \in R^3$ in its quaternion representation with the filter $F \in R^2 : [Mc = I * F, \in R^2]$, where the value of each pixel of Mc is obtained by using the discrete convolution (Equation 7), using a filter of $(n \times m \times 4)$ as follows:

$$Mc(x,y) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{4} I(i,j,k)\, f(i,j). \tag{7}$$

In Figure 7, the colors that seem to highlight more are marked in the RGB triangle with a white cross, these are however, the ones that are far away from the background color in the triangle (indicated with a circle surrounding the white cross). Figure 8 shows the contrast map implemented with a magnitude of the gradient filter. In the three cases the circles that highlight the most are the ones indicated with the white cross, and the ones that highlight least are those near the background color, marked with black cross. What we can infer from the saliency of the contrast can be in function of the absolute difference

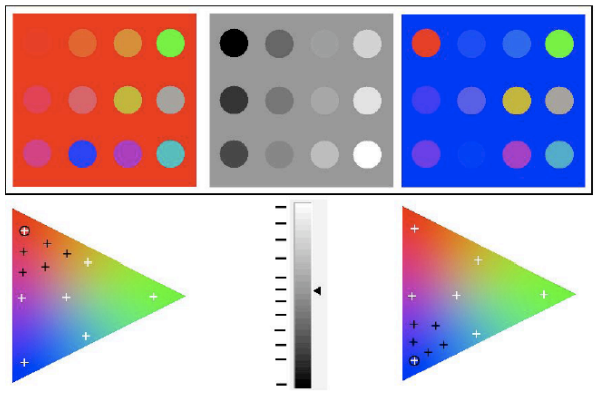


Figure 7: Color and grayscale images where different contrast can be appreciated. The background color is represented in the RGB triangle with a cross and a circle while in the grayscale bar with an arrow; the rest of the colors are marked with a cross or small lines, respectively. It can be observed in the color images that the white crosses highlight the colors that are far away from the background color.

between pixel and its neighbors, either vectorial for colors or scalar for intensities. This is the reason we use a magnitude of the gradient filter, besides of being faster than a high-pass filter.

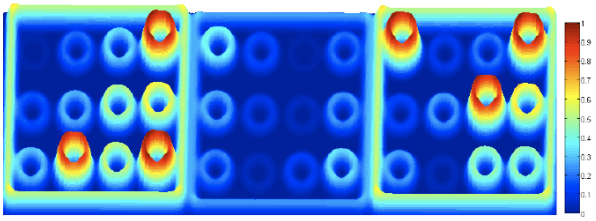


Figure 8: Color and intensity contrast maps of the images in Fig. 7 using a magnitude of the gradient filter. It can be seen that the circles that highlight the most are the ones indicated with the white cross. The saliency of the contrast seems to be a function of the absolute difference between pixel and its neighbors.

We now show the contrast maps of a scene full of high contrasts in Figure 9, where the regions of maximum contrast are in red.

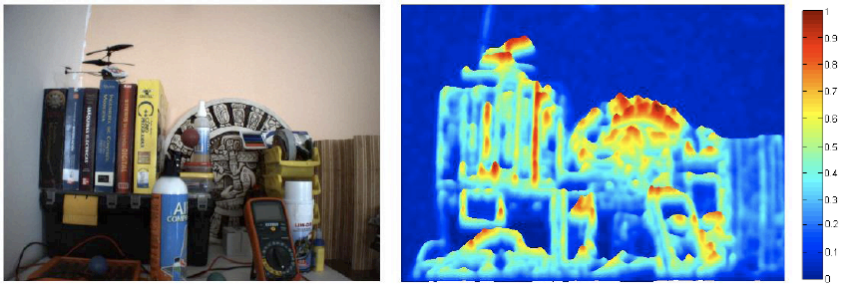


Figure 9: Contrast and color map of a scene full of high contrasts. Left image shows the input color image and the right image the contrast map. It can be observed that the regions of maximum contrast are in red.

5.2 Motion map

The individual motion of pixels in an image is known as optical flow and is measured when corresponding pixels are found in two consequent frames. On one hand, the measurement of the optical flow can be highly complex due to the similarity between pixels, and with current hardware technology, is almost impossible to implement this algorithm on a real-time system. On the other hand, the main attentional map is the motion map as is related with the most important task of the system: tracking moving objects. An easy way to obtain a motion map without solving the correspondence problem is by using a motion filter of absolute rate [3]. This filter consists on subtracting two images (I) taken at different instants of time $(t, t-1)$ and observing the regions for which the squared difference is maximized. The absolute rate of motion is given by $M_t = I_t - I_{t-1}$. The result can be seen in Figure 10. M_t can operate as a motion map, however when convoluting a mean filter on M_t, we obtain a more precise reading of the region with the greatest motion.

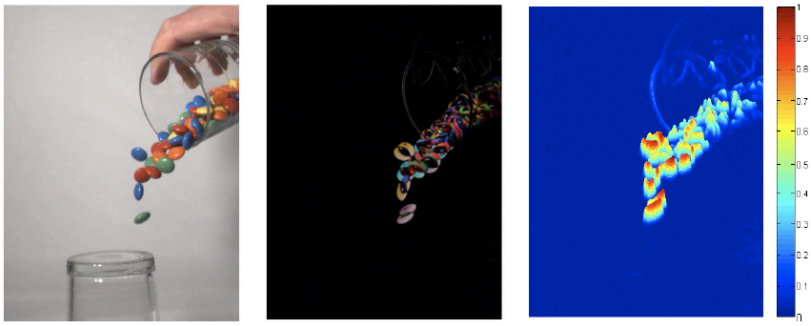


Figure 10: Motion absolute rate of a scene. The left image is the original image, the middle image is the motion absolute rate and the right image is the normalized motion map, where the regions that change the most are those that attract more the attention.

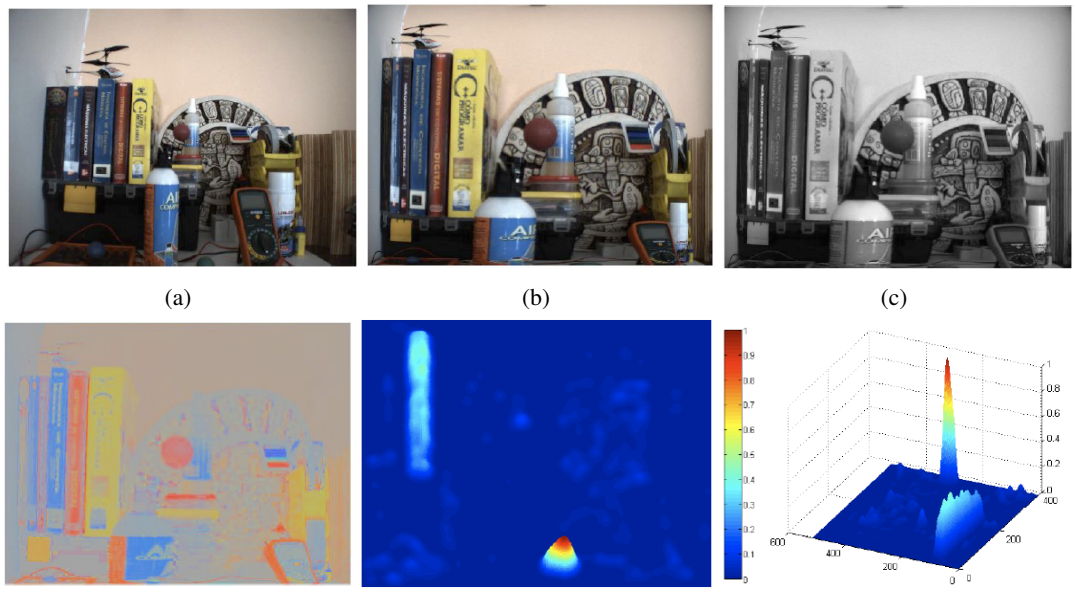
Obtaining this map could be complicated when the vision system is active, as the motion of the camera generates a visual flow in the whole image, for this reason the use of this map will be done exclusively at the end of each saccadic motion, when the motion of the cameras is practically null.

5.3 Inhibition of return

Inhibition of return is one of the most important components in the visual attention process as it retains in memory the regions that were already attended, to incentive the recollection of information in regions that have not been attended. This inhibitory effects does not depend on visual cues but on their spatial locations, for which we store the position of fixation points and induce an artificial potential field built in the robot's workspace, so the inhibition map is the projection of the sum of the potential fields on the image. We define the potential field on a point as:

$$U(\theta) = 1 - \sum_{i=1}^{n} e^{\frac{\theta_i^2}{\sigma^2}}, \tag{8}$$

where θ is the angle between vectors A and B, σ is the constant used to restrict the dilatation of the potential field so that does not exceed a solid angle greater than that occupied by the fovea. And, in order to project the effect of the potential field to an inhibition of return map (IoR), we use the coordinates (u,v) of vector B: $IoR(u,v) = U(\theta)$.



(a) (b) (c)

(d) (e) (f)

Figure 11: Saliency map based on the red sphere. (a-b) are the original and foveated images, respectively; (c-d) quaternion representation of (b); (e) histogram map slightly rotated on the horizontal axis; (f) the same map inclined and rotated on the Z axis.

5.4 Map integration

Under this computational model, visual attention is seen as a cost function that varies over time. Each visual characteristic constitutes a variable that, depending on its manipulation, can make fluctuate the attentional focus of the system giving an appearance of an animal or human behavior before an unknown scene. The integration of all maps is the more difficult and challenging part of the whole design. In our system we assign the preferences (weights) to each map according to the priority of each of the tasks.

We show in Figure 11 an experimental result of the saliency map obtained after integrating all the maps described above. In the sequence of images we show a scene where the object of interest is the red sphere. It can be seen that the sphere gets far from the fovea (located first at the center of the image) and therefore we used an attention map based on the color of the sphere to find it. Despite that there exist more red objects in the scene, the histogram resolution is good enough to broadly distinguish the sphere over the other objects. Even though this map is enough for this case, there will be other cases in which there exist very similar objects and the difference is not so well marked. Those cases can be solved using information about the motion of the cameras in order to see the regions of interest in high resolution, obtain fine details and facilitate segmentation.

6 Conclusions and Future Work

We have presented a visual attention algorithm that combines existing approaches for segmenting moving objects in real time. First, the proposed image compression technique (foveated vision) with multiple resolution allows us to process images four to seven times faster and accomplish the objective of scanning the complete field of view without having to use special cameras. Second, the use of a quaternion representation of the RGB space together with the color histograms allow us to identify objects in a robust manner before illumination changes; so, we can use the same algorithm during the day with ambient illumination and during the night with artificial light. One clear disadvantage, however, about

using color is the increase in the processing time. This forces us to select carefully the maps to be integrated and reduce its number to three. We hope in the near future to improve the latencies of the processing algorithms to include other maps, such as depth maps to direct the sight to areas where no measure of depth has been done.

6.1 Acknowledgments

The authors thank the National Council of Science and Technology (CONACyT) for funding this project.

References

[1] B. T. Backus, M. S. Banks, and J. A. Crowell R. van E. Horizontal and vertical disparity, eye position, and stereoscopic slant perception. *Vision Research*, 39:1143–1170, 1999.

[2] M. Bjorkman and D. Kragic. Active 3d scene segmentation and detection of unknown objects. In *International Conference on Robotics and Automation (ICRA)*, 2010.

[3] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati. Active vision for sociable robots. *IEEE Trans. on Systems, Man, and Cybernetics,*, 31(5), 2001.

[4] B. Espiau, F. Chaumette, and P. Rives. A new approach to visual servoing in robotics. *IEEE Transactions on Robotics and Automation*, 8(3), 1992.

[5] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2, 2001.

[6] W. James. *The Principles of Psychology*, volume 1,2. Harvard University Press, 1981.

[7] R. M. Klein. Inhibition of return. *Trends Cognitive Science*, 4:138–147, 2000.

[8] C. Koch and S. Ullman. Selecting one among the many: a simple network implementing shifts in selective visual attention. In *Artificial Intelligence Lab Memo No. 770.* MIT, Cambridge, 1984.

[9] A. Mishra, Y. Aloimonos, and C. L. Fah. Active segmentation with fixation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

[10] A. Mishra, Y. Aloimonos, and C. Fermuller. Active segmentation for robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3133–3139, 2009.

[11] K. Nakayama and M. Mackeben. Sustained and transient components of visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, 14:453–471, 1989.

[12] N. Oshiro, N. Maru, A. Nishikawa, and F. Miyazaki. Binocular tracking using log polar mapping. In *IEEE/RSJ Intl. Conference on Intelligent Robots and Systems*, pages 791–798, 1996.

[13] Pahlavan and Eklundh. A head-eye system - analysis and design. In *CVGIP: Image Understanding: Special issue on purposive, qualitative and active vision*, 1992.

[14] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *Int. Journal of Computer Vision*, 18:233–254, 1996.

[15] E. L. Schwartz. Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics*, 25(4):181–194, 1977.

[16] D. L. Sparks. The brainstem control of saccadic eye movements. *Nature Rev. Neuroscience*, 3, 2002.

[17] H. V von Helmholtz. Treatise on physicological optics. *Optical Society of America*, 3, 1925.

[18] H. Wallach and L. Floor. The use of size matching to demonstrate the effectiveness of accomodation and convergence as cues for distance. *Percept. Psychophys*, 10:423–428, 1971.

[19] G. Welch and G. Bishop. An introduction to the kalman filter. In *Tech. Report TR 95-041*. University of North Carolina, Department of Computer Science, 1995.