Evaluation of Visual Attention Models for Robots

M. Zaheer Aziz, Bärbel Mertsching, M. Salah E.-N. Shafik, Ralf Stemmer
GET Lab, University of Paderborn, Paderborn, Germany
{aziz, mertsching, shafik, stemmer}@upb.de

Abstract

This paper presents a new approach for providing visual attention on robot vision systems. Compared to other approaches our method is very fast as it processes regions rather than individual pixels. The proposed method first builds a list of regions by applying a shade and shadow tolerant segmentation step. The features of these regions are computed using their convex hulls in order to simplify and accelerate the processing. Feature values are stored within the records of respective regions instead of constructing a master map of attention. Then an algorithmic method is applied for finding the focus of attention in contrast to mathematical approaches used by existing models. Experiments conducted on simulated and real image data have not only demonstrated the validity of the proposed approach but have also led to the establishment of a comprehensive robotic vision system.

1. Introduction

Mobile robot vision systems require simple and efficient algorithms as they carry limited computational resources in order to reduce energy requirements and construction cost. Visual attention aims to mimic the ability of natural vision systems [6] to select just the relevant aspects from the visual input. This helps reducing the processing time needed for high level object recognition tasks for given input images. Hence attention can play a vital role in making vision systems work in real time. Generally, an attention algorithm starts with the computation of saliencies in an image with respect to different features such as symmetry, eccentricity, color contrast, and orientation, etc. (see e.g. [3] and [14]). Different saliency maps for these features are generated which are finally combined into a master map of attention. In the next step, a mechanism for selecting the focus of attention (FOA) is applied. In order to accomplish a continuous operation, an inhibition of return (IOR) is incorporated whose responsibility is to provide equal opportunity to all objects and not to keep attending a single object while ignoring others.

In psychophysics, two types of attention behaviors are distinguished: while in covert attention salient portions of the input are attended internally without making any sensor movements, overt attention is performed by moving the sensor towards the attended object. Hence for the later, it is not only important to have an attention module in a robot system, but vice versa for the verification of an attention model, a successfully running robot vision system is required.

One of our major objectives is to simply artificial visual attention so that the algorithm becomes easily transferable and executable on limited hardware resources. Our earlier work was based on the common approach while now we concentrate on a region-based method. The objective is to obtain a FOA not amid an activity cluster but amid the shape of an actual object. This approach allows the reuse of the selected regions of interest in the further steps, e.g. object recognition, instead of requiring to start a redundant process for this purpose after completion of the attention procedure.

2. Existing Models of Visual Attention

Artificial visual attention has been a topic of interest for many computer vision applications. For example, it was used in autonomous vehicles with ability to maneuver intelligently using regions of attention to determine optimal gaze for multi-focal saccadic vision [11]. Another interesting application is presented in [12] that shortlists objects in a given scene to find out an expected target object pointed by hand or finger in a wearable gesture recognition system. Visual attention algorithms have been incorporated in image compression techniques such as JPEG 2000 in order to improve view quality of important objects in compressed images [7].

The pioneering work on artificial visual attention as presented in [13], [14], and [15] builds saliency maps for three features, namely, color channels, intensity, and orientations. Each feature is computed by a set of linear center-surround operations between fine and coarse scales analogous to visual receptive fields.
These feature maps are combined into three conspicuity maps for intensity, color, and orientation through across-scale addition. At any given time, the maximum of sum of saliency maps defines the most significant image location to which the focus of attention should be directed. This is done by a 2D layer of leaky integrate-and-fire neurons. This layer feeds into a biologically plausible winner-take-all (WTA) neural network. Shift of attention to the winner location causes a global inhibition of all WTA neurons and transient activation of local inhibition.

The method described above has been an inspiration for many other models for visual attention. The model presented in [18] and [19] introduces new feature maps for edges and symmetry. It computes two color maps in contrast to one in the previously mentioned models. This model uses a Gaussian pyramid with different scales from 0 to n levels, where each level is generated by sub-sampling of $2^n$ for constructing the four feature maps. The center-surround is implemented as the difference between fine and coarse scales of Gaussian pyramid images. In total 24 maps are computed and combined into four conspicuity maps. Unsupervised learning is used to determine the relative importance on different bases to generate a suitable salient region. The IOR process is implemented by masking off the currently attended focus of attention for the next attention cycle.

The model presented in [21] computes object based saliency depending on groupings. A grouping is considered as a hierarchical structure of “objects and space”; hence it may be a point, an object, a region, or a structure of other groupings. The primary features are extracted exactly as done in [15] but it constructs the intensity, color (red, green, blue, and yellow), and orientation pyramids after applying a Gaussian filter $W_{g}$ and a Gabor steerable filter $W_{gs}(\lambda; \theta)$ on the five feature channels of intensity, red, green, blue, and yellow. In this model the shift of attention is done by using an algorithmic approach with a coarse to fine strategy. Another work as presented in [16] uses a fuzzy growing method to extract attention windows after applying a local contrast technique for saliency map generation.

Our earlier model presented in [3] and [4] computes four types of feature maps namely, symmetry, eccentricity, color contrast, and depth. For symmetry, Gabor filters in twelve orientations are applied and results are added pixel wise for all orientations and stored in a so-called radius image. This is done separately for different radii with adjoining addition intervals. The maximum value for a set of different radii around each pixel is chosen as the saliency value of symmetry for. The eccentricity feature is obtained by principle component analysis of segments in the input. For computing color contrast feature, saliency is computed in the polar form of MTM color space. In order to compute depth feature, Gabor filters with vertical components are used on both left and right stereo frames. The FOA is chosen in a two stage mechanism: First the weighted sum of feature maps is fed to a system of neural fields that performs an all-to-some selection which results in a small set of activity clusters as output. Then in the second stage, an inhibition of return process selects one activity cluster at a time for focusing the attention according to required behavior.

The computation time of some of these methods is too high making them unable to operate on real time vision systems. Even the fastest one would take above a second to process a complete image. Working on individual pixels or a small neighborhood not only causes redundant addressing of a single point but also leads to activity clusters for FOA in cloudy shapes. Hence a region based approach is proposed to achieve focus of attention in shape of actual objects and to minimize redundant processing.

3. Proposed Method

Keeping in mind the objectives mentioned in the first section, the proposed method works on a simplified algorithmic approach for feature extraction and selection of attention foci. Some of the procedures that are treated independently in the existing methods are integrated together in order to reduce processing cycles. The basic idea is to confine the feature processing and computation of FOA to pixel groups or segments in the input. Segments represent natural groupings of pixels that belong together and most of the times they correspond to complete objects in a scene. Hence our approach first addresses to color segmentation [10] of the visual input. The resulting segments are later used for computing features via their convex hulls. In the current status of proposed approach, four features including color contrast, eccentricity, orientation, and symmetry are considered. More feature channels may be added in future work.

Computation of color contrast is coupled with the segmentation step in order to complete both steps in one cycle. The HIS (Hue-Intensity-Saturation) color space is considered as an appropriate representation of human color perception [9]. Hence we use this space in our model for segmentation to reduce the influence of illumination changes and shadows on extracted regions [2]. Seeds with color saturation above a high threshold $c^*$ are selected for a region growing process. Hence, the set of seed pixels $S^*$ from an image input $I$ can be defined as:
constructed as follows:

\[ S' = \{ P(x, y) \mid \text{sat}(P) > \tau, \forall (x, y) \in I \} \] (1)

where sat\((P)\) gives saturation of the pixel color. The color of seed pixel decides the way in which the region will be grown further. Hence the list of regions \( R_i \) is constructed as follows:

\[
P(x, y) \mid \begin{align*}
    \text{int}(P) < \tau & \quad \forall (x, y) \in NH(S_i') \\
    \text{int}(P) > \tau & \quad \forall (x, y) \in NH(S_i') \\
    \text{sat}(P) < \tau & \quad \forall (x, y) \in NH(S_i') \\
    \text{sat}(P) > \tau & \quad \forall (x, y) \in NH(S_i') \\
    \text{hue}(P) < \tau & \quad \forall (x, y) \in NH(S_i') \\
    \text{hue}(P) > \tau & \quad \forall (x, y) \in NH(S_i')
\end{align*}
\] (2)

where \( \text{int}(P), \text{hue}(P), \) and \( \text{sat}(P) \) describe the intensity, hue, and saturation component of the color of pixel \( P \) respectively. \( NH(S_i') \) denotes the neighborhood of seed \( S_i' \), whereas \( \tau, \tau_i, \tau_d, \tau_i \) and \( \tau_d \) are the thresholds for low intensity, high intensity, low saturation, hue difference, and relaxed saturation respectively.

Too small regions are deleted before proceeding further because they are expected to be result from noise. The remaining unprocessed area is evaluated after segmenting the image completely for high contrast seeds. If a large area of image remained unprocessed then threshold for color contrast \( \tau \) is lowered and the remaining portions of the image are segmented with this relaxed threshold. The final output of this process is a list of regions loaded with information about their location, color contrast, and vertical bounding rectangle. The saturation component of color of seed pixel is saved as color saliency \( C_i \) for every region \( R_i \).

Further features are computed based upon the convex hulls of the acquired regions. In general the convex hull for a finite set of points \( P \) is defined as the smallest convex subset of \( P \) that contains \( P \) [17]. The resulting set can be imagined as a group of points that would lie on a stretched rubber band around the given set \( P \). We extract convex polygons \( \phi_i \) for each region \( R_i \) obtained in the segmentation process using an efficient outer scan method as proposed in [1].

As our approach works on complete regions, we compute individual saliency values for eccentricity rather than comparing the neighborhood. The computation of the eccentricity feature is done by calculating the following distances for each \( \phi_i \), where \( \text{dist}(V_i, V_j) \) gives distance between polygon vertices \( V_i \) and \( V_j \):

\[ d_i = \max(\text{dist}(V_i, V_j)) \quad i \in [1..n-1], \quad \text{for each } i: j \in [i+1..n], \]

\[ d_2 = \max(\text{dist}(mid(V_i, V_{i+1}), mid(V_j, V_{j+1}))), \quad i \in [0..n-1], \quad \text{for each } i: j \in [0..n], \]

\[ d_3 = \max(\text{dist}(mid(V_i, mid(V_{i+1}))), \quad i \in [0..n], \quad \text{for each } i: j \in [0..n], \]

\[ D'_{maj} = \max(d_1, d_2, d_3) \] (3)

Where \( D'_{maj} \) is length of the major axis \( M_i \) of \( \phi_i \). Now, we define \( P_R \) as the set of vertices at right side of \( M \) and \( P_L \) as set of vertices at left side of \( M \). The length of minor axis (width) of each \( \phi_i \) is computed in two parts \( D'^{Ri}_{min} \) and \( D'^{Lj}_{min} \) which are the extents at the right and left side of \( M \) respectively:

\[ D'^{Ri}_{min} = \max(\text{dist}_{lp}(R_i, M_i) \forall R_i \in P_R), \]

\[ D'^{Lj}_{min} = \max(\text{dist}_{lp}(L_j, M_j) \forall L_j \in P_L) \] (4)

where the function \( \text{dist}_{lp}() \) gives point to line distance. Now eccentricity feature \( E_i \) of every region \( R_i \) can be simply computed as:

\[ E_i = (D'^{Ri}_{min} + D'^{Lj}_{min}) / D'_{maj} \quad \text{where } i \in [1..n] \] (5)

The saliency of orientation is by nature a feature relative to neighbourhood of an object. In this paper we discuss orientation saliency relative to all regions in a given input frame. Let \( \theta \) be the inclination of \( M_i \) of \( \phi_i \) then we define a counting function \( T(\theta, \theta) \) as:

\[ T(\theta, \theta_j) = \begin{cases} 
    0 & \text{for } |\theta_j - \theta| \leq t \forall j \in [1..n], j \neq i \\
    1 & \text{otherwise}
\end{cases} \] (6)

where \( t \) is a small threshold value. Now orientation saliency of \( \phi_i \) is computed as:

\[ O_i = \sum_{j=1}^{n} T(\theta, \theta_j) \] (7)

For evaluating the symmetry of a region \( R_i \) we first find its center of gravity \( C_x^i \) by:

\[ C_x^i(x, y) = \sum_{j=1}^{n} x_j / n, \quad \sum_{j=1}^{n} y_j / n \quad \forall (x, y) \in \phi_i \] (8)

Now if \( \rho \) denote the set of distances of all points of \( \phi_i \) from \( C_x^i \) then we take \( S_1^i \) as first step of symmetry. Let \( \text{dev}(\rho) \) be a function giving the standard deviation of given \( \rho, t \), \( t \) be a threshold, and \( \delta \) be a predefined constant, then \( S_1^i \) for each region is defined as:

\[ S_1^i = \begin{cases} 
    \delta \text{ when } \text{dev}(\rho_i) < t \delta \\
    0 \text{ otherwise}
\end{cases} \] (9)

Let \( \gamma \) be the highest possible saliency value for symmetry (here \( \gamma = 255 \)) then second step of symmetry for each \( \phi_i \) is calculated using values from equation (4) as:

\[ S_2^i = \gamma / 3 - (D'^{Ri}_{min} - D'^{Lj}_{min}) \] (10)
For third step of symmetry, center of gravity for $P_L$ and $P_R$ are computed into $C_L^t$ and $C_R^t$ and one end of $M$ is taken as reference point $m_r$ for each $\phi_i$. Now the step three symmetry $S_i^t$ for each region is computed as:

$$S_i^t = \gamma 3 - \left| \text{dist}(C_i^t, m_r) - \text{dist}(C_i^t, m_l) \right|$$  \hspace{1cm} (11)

Now the final saliency value $S_i$ for symmetry of $\phi_i$ is the sum of $S_i^t$, $S_i^t$, and $S_i^t$. 

The proposed approach for selection of FOA takes its influence from the model given in [3], in which weights for different feature maps are dynamically altered in the iterations of IOR process. The weighted sum of features is stored in a master map which is in turn taken by a system of neural fields as input. We propose an algorithmic approach functionally equivalent to a neural one, working directly on feature maps instead of forming a master map. Let $F_i$ be the set of features for the region $R_i$ such that:

$$F_i = \{ C_i, E_i, O_i, S_i \} \text{ for } i \in [1..n]$$

Now we define a function $\phi(F_i)$ that picks a single feature from each $F_i$ at time $t$, for example $\phi(F_i) = C_i$ at $t = 1$. This function is cyclic therefore $\phi^{t+4}(F_i) = \phi(F_i)$. Let $f$ be a feature value such that

$$f = \max \left( \phi(F_i) \text{ for } i \in [1..n] \right)$$

then focus of attention at a time $t$ will be a $R_i$ such that $\phi(F_i) = f$. After attending $R_i$, its features are suppressed by a predefined value $\varepsilon$ for inhibition of return as:

$$\phi^{t+4}(F_i) = \phi(F_i) - \varepsilon$$

4. Experimental Platforms

A robust and comprehensive testing platform is necessary to test the validity of developed algorithms before running them in a real-life environment. For the case of attention algorithms, covert attention can be tested with static images as input to simple computer programs but testing of overt attention requires a more sophisticated infra-structure of an entire vision system. There are two ways to establish a test environment for experiments. It is equipped with a stereo camera head and a wide range of sensors like electronic compass, GPS device, infrared sensor, and motion detector. It can be remotely controlled by input devices such as joystick, force-feedback steering wheel, and head mounted display with head-tracker. This provides a very intuitive man-machine interaction for tele-presence applications. Operator/host and tele-operator communicate via wireless links. Due its multitude of capabilities, it can also be used for testing of algorithms and models for autonomous control and maneuvering. A glimpse of this mobile system is given in figure 1 with a closer look at the camera head in the top-right corner.

5. Experiments and Results

The proposed model of visual attention has been tested on both test platforms mentioned in section 4. For describing the general performance of our
approach, we present some experiments and results using the TSR. A simple scene as shown in figure 2 was set up in front of the robot. Results of feature saliency computations are presented as gray-scale images in the same figure. The brighter the color the more salient an object is in the particular feature map.

Figure 3 shows feature bars at four time steps: In the left top diagram (first frame), the focus of attention is determined on color contrast and the yellow ball (Yb) yielding the highest contrast is attended which is indicated by a small vertical line below the bar. Then Yb is suppressed using the IOR mechanism. Now more emphasis is put on the symmetry feature. The bottle (Bo), blue ball (Bb), and red ball (Rb) are attended using the same mechanism.

The required gaze shifts are transformed into motion commands forcing the camera head to focus on the object to be attended, which means that the object is centered in the frame. Figure 4 demonstrates overt attention of the vision system according to the sequence of FOAs selected according to the results given in figure 3.

Figure 2. Input scene to the vision system, Eccentricity map, Orientation map, Color Contrast map, Symmetry map.

Two virtual images from our simulator and two real images have been used to compare the proposed model with the model of Itti and Koch [15] and our previous work [3] in different contexts. The results are summarized in figure 5. The top row shows results by using the model of Itti and Koch, the middle row presents results of our previous system, and the results of proposed model are given in the bottom row. It can be observed that Itti’s model misses the ball (an important item) in the second column. Both of the existing methods ignore the car in the traffic scene given in the rightmost column.

Another significant advantage of the proposed scheme is its efficient computation time. Figure 6 depicts the comparison of processing time between the existing model of [15] and the proposed one. Computation time for the model of [3] was too high to be adjusted to the comparison graph. Our model is approximately 500 times faster than the model of Itti.

6. Discussion

The outcome of the research presented in this paper is two-fold. Firstly, an attention model has been proposed that is strong in performance but low in computational complexity and can therefore be used for real time applications in natural scenes on real robot platforms. Although the results obtained are very promising the proposed technique it is still in an early
stage of development. More work is obviously required to enhance the level of robustness. Secondly, a complete vision system for evaluation of models of visual attention has been established.

Acknowledgement

We thank Dirk Fischer and Lars Heeper for their assistance in implementing the proposed attention model on the TSR.

Fig. 5. Comparison of CPU time (in log scale) between proposed method and one of the existing ones.

References