Visual Attention
Using Game Theory

Ola Ramström and Henrik I Christensen

Computational Vision and Active Perception
Numerical Analysis and Computer Science
Royal Institute of Technology
SE-100 44 Stockholm, Sweden
olar@cvap.kth.se
bic@cvap.kth.se
http://www.biom.kth.se/

Abstract. A system using visual information to interact with its environment, e.g. a robot, needs to process an enormous amount of data. To ensure that the visual process has tractable complexity visual attention plays an important role. A visual process will always have a number of implicit and explicit tasks that defines its purpose. The present document discusses attention mechanisms for selection of visual input to respond to the current set of tasks. To provide a truly distributed approach to attention it is suggested to model the control using game theory, in particular coalition games.

1 Introduction

The amount of visual information available to a system is enormous and in general it is computationally impossible to process all of this information bottom-up [10]. At the same time it is essential to note that a vision system is of limited interest when considered in isolation. A system will always have a set of tasks that defines the purpose of the visual process. To ensure that the process has tractable computational properties visual attention plays a crucial role in terms of selection of visual information. Visual attention is divided into two different components: overt attention – the control of the gaze direction and covert attention the selection of features, and internal control of processes. The present document only discusses covert attention.

The literature contains a number of different models for visual attention, a good overview of models can be found in [6]. An important goal of visual attention is to focus on the most important information given the systems current goals/tasks. To accommodate this the system must select the most relevant information for further processing. An obvious question is: how can the system choose the most important information without prior processing of all information to determine the utility of different cues. If attention is engaged early in the visual chain it is necessary to consider how models of objects, situations
and tasks can be utilized to “predict” the value of different cues (property selection) and the approximate location of such cues (spatial selection). The literature contains at least three major theories of visual attention: i) The Spotlight Metaphor [7] models selection as highlighting of a particular region while other regions are suppressed, i.e., a particular region or set of cues stand out. Once a region has been processed the spotlight can be moved to a new area through a disengage/move/engage process. This model has also been extensively studied computationally by [4] and [11]. ii) The Zoom Lens metaphor [2], where a zoom type mechanism is used for selection of the scale (the size of the field of processing), i.e., the “natural” scale is selected and one cannot attend to a leaf and the overall tree structure at the same time. iii) Object based approaches where selection is closely tied to the object/event of interest. While i) and ii) are directly related to selection of spatial regions iii) is tied to objects, parts, context, or relations. Object based approaches have been reported by [1], and evidence in association with tracking has been reported by [8]. For all three approaches it is essential to consider the mechanism for selection and its relation to cost, capacity and task/object information.

To accommodate intelligent spatial selection it is of interest to study methods that allow distributed processing of visual information to allow efficient control of the process with a minimum of centralized coordination. Centralized coordination would impose a challenge in terms of biological plausibility and it would at the same time introduce a potential bottleneck into a system. Distributed mechanisms for control are thus of interest, irrespective of the chosen attention model. One area that has considered distributed control extensively is game theory [3]. Both differential games and coalition games have interesting theoretical models that potentially could be used to study coordination/control mechanisms. This paper provides a discussion of the potential use of such a game theoretical model for visual attention.

1.1 Overview

The document is focused on a discussion of saliency measure using multiple cues and game theory concepts. As a precursor to this the set of feature maps used for the analysis are outlined in section 2. These feature maps are integrated using a scale space pyramid, the method is here similar to the attention mechanism described by Tsotsos in [11]. The integration of features and processing of the pyramid is outlined in section 3. The nodes of the pyramid are subject to trading on a market, the outcome of the trading represents the saliency. The basic dynamics of the market is outlined in section 4.1 and the implementation of the market in the pyramid is outlined in section 4.2. In section 5 we discuss the spotlight mechanism for finding the region of interest. Finally a set of experiments are described in section 6.
2 Feature Maps

The focus of the paper is on distributed selection of multiple cues for recognition with emphasis on selection and control. Consequently a relatively simple set of features has been selected to demonstrate the concept. The feature detectors operate on a color image of size 720*576 (see image 4 top left). For the experiments we chose to use the following 4 feature detectors:

1. Pseudo red (red): the red channel divided with the intensity (R/(R+G+B)).
2. Pseudo green (green): similar to the pseudo red.
3. Pseudo blue (blue): similar to the pseudo red.
4. Intensity (white): the intensity (R+G+B).

For more complex images it is obvious that more complex features/cues will be needed, but that is not the real issue here.

3 The Pyramid

We will use a pyramid for multi-scale attention processing. The pyramid is similar to the one presented by Tsotsos [11]. The responses from the feature maps are input to the pyramid. Each node in the pyramid will sum the values from a subset of the nodes in the layer below, see figure 1.

![Pyramid Diagram]

Fig. 1. One dimensional example of the integration across scales for one feature.

Since all layers in the pyramid have the same size, we can combine all feature maps and view each node as a vector of all features, see figure 2.

As we propagate the feature values up in the pyramid, a node in layer $l+1$ that is connected to the set $A_l$ in layer $l$ will get the value:

\[ n_{x,y,l+1} = \sum_{x,y \in A_l} n_{x,y,l} \]
where $n \in \mathbb{R}^k$ and $k$ is the number of features maps. A vector $n$ corresponds to a feature dimension as discussed in [9].

Let us consider an example where we have one feature map for the red channel (R), one for the green channel (G), and one for the blue channel (B) then the node at position $(x, y)$ in layer $l$ has the value

$$n_{x,y,l} = (R_{x,y,l}, G_{x,y,l}, B_{x,y,l}).$$

A task for the attention system is described as a wanted feature vector $(w)$, e.g.,

$$w = (R, G, B).$$

Similarity is measured as:

$$\frac{w \cdot n_{x,y,l}^T}{|w||n_{x,y,l}|}$$

Note that the absolute value of $n$ and $w$ is not interesting here. If we were interested in $|n| = \sqrt{R^2 + G^2 + B^2}$, i.e. brightness (I), we would add that feature to $n$ and $w$:  

$$n_{x,y,l} = (R_{x,y,l}, G_{x,y,l}, B_{x,y,l}, I_{x,y,l})$$

and

$$w = (R, G, B, I).$$

Let us define $n_{A_l}$ as the sum

$$n_{A_l} = \sum_{x,y \in A_l} n_{x,y,l}.$$

Let $A'_l$ be a subset of $A_l$, see figure 2. Then we can define a salient region $A_l$ as

$$\frac{w \cdot n_{A_l}^T}{|w||n_{A_l}|} > \frac{w \cdot n_{A'_l}^T}{|w||n_{A'_l}|}$$  \hspace{1cm} (1)
That is, a region that matches the wanted feature vector better than its surrounding. In section 4.2 we will discuss how we calculate and compare saliency. The most salient node (at any scale) is selected. From that node we form a spotlight that points out the attention region in the image, In section 5 we will discuss how to that is done.

Finally, the selected region is inhibited and a new search can be carried out.

4 Saliency Computing

4.1 A Market

Competitive equilibrium of a market is commonly used in classical economy. A market is a place where actors can buy and sell goods. With a set of goods each actor can produce a value, The produced value is denoted utility, and the function that maps a set of goods to a utility value is concave. If the utility of the market can be shared among its members in arbitrary ways, e.g. if we use money exchange, we say that the market has a transferable payoff. In [5] it is shown that a market with transferable payoff will reach a competitive equilibrium. A competitive equilibrium is a state where all actors agrees on a price for each type of good and the actors will buy and sell until all have the same amount of all types of goods.

Let us consider a market with \( N \) actors and \( k \) number of available goods, Actor \( i \) has an allocation of goods \( n_i \in \mathbb{R}^k \) and the utility \( f(n_i) \in \mathbb{R} \). The utility function is concave and therefore most actors will gain on selling and buying goods on the market.

After trading actor \( i \) will have a new allocation of goods \( z_i \), where \( \sum_{i\in N} z_i = \sum_{i\in N} n_i \). Each agent will strive to get the \( z_i \) that solves

\[
max_{z_i} \left( f(z_i) - p \cdot (z_i - n_i)^T \right) \tag{2}
\]

where \( p \in \mathbb{R}^k \) a prize vector.

We denote the average allocation \( \bar{n} = \sum_{i\in N} n_i / N \). In [5] section 13.4 it is shown that the solution where \( z_i = \bar{n} \) for all \( i \in N \) and the price vector \( p = f'(\bar{n}) \) is a competitive equilibrium.

4.2 The Feature Market

Goods that are rare on a market and that increases the utility of the actors are expensive. If the inequality in equation (1) is large then the nodes in \( A' \) will sell at a high prize to the nodes outside \( A' \). The wealth of the region \( A' \) depends on the value of its features and the need for them outside \( A' \). We will define a wealthy region as a salient one.

In the proposed solution we use a market with transferable payoff (c.f. section 4.1) to define saliency.

In this market the goods are features and the actors are nodes. Let us define:
\(- \( k \) is the number of features,
- \( l \) is a layer in the pyramid.
- \( A_l \) is a set of nodes in layer \( l \).
- \( A_l' \) is a subset of \( A_l \).
- \( A_l \setminus A_l' \) is the set of nodes \( A_l \) excluding \( A_l' \).
- \( n_i \in \mathbb{R}^k \) is the measured feature values at node \( i \in A_l \).
- \( n_i' \in \mathbb{R}^k \) is the measured feature values at node \( i \in A_l' \).
- \( \hat{n} = \sum_{i \in A_l} n_i / |A_l| \).
- \( \hat{z}_i \in \mathbb{R}^k \) is the allocation of feature values after trading at node \( i \in A_l \).
- \( f(z_i) = w \ast z_i^T / |z_i| \) is the utility of node \( i \), where \( w \in \mathbb{R}^k \) is the wanted allocation given by the task.
- \( p = f'(\hat{n}) = \frac{w}{|\hat{n}|} - \hat{n} \ast \frac{w \hat{n}^T}{|\hat{n}|} \) is a prize vector.

In section 4.1 we saw that \( f(z_i) \) have to be concave. We observe that the \( f(x) \) used in this solution is concave:

\[
\frac{f(a) + f(b)}{2} = \frac{w}{2} \ast \frac{a^T + b^T}{|a + b|} \leq w \ast \sum_{i \in A_l} n_i / |A_l| = f\left( \frac{a + b}{2} \right)
\]

In section 4.1 it is shown that the actors will trade goods on the market to get the allocation \( \hat{n} \) that solves equation (2).

When the market has reached its competitive equilibrium all nodes will have the allocation \( \hat{n} \). The nodes that have sold more features than they have bought have a positive capital. From equation (2) we can derive that the capital \( C \) of actor \( i \) is

\[
C_i = p \ast (n_i - \hat{n})^T = w \ast (I - \hat{n}^T \hat{n}) \ast (n_i - \hat{n})^T;
\]

\[
|\hat{n}| = |n_i| = 1.
\]

The saliency of node \( n_{t+1,x,y} \) which is connected to the set \( A_l \) is

\[
S = p \ast (n_i' - \hat{n})^T = w \ast (I - \hat{n}^T \hat{n}) \ast (n_i' - \hat{n})^T;
\]

\[
|\hat{n}| = |n_i'| = 1.
\]

Thus, the saliency of the node \( n_{t+1,x,y} \) is the capital achieved by the connected nodes \( A_l' \).

4.3 The Context Matrix

The saliency of node \( n_{t+1,x,y} \) is defined in equation (3). We can rewrite that equation as:

\[
S = w \ast A \ast b^T;
\]

where \( A = I - \hat{n}^T \hat{n} \), \( b = n_i' - \hat{n} \), and \( |\hat{n}| = |n_i'| = 1 \).

The vector \( b \) has the expected properties that the more the feature vector differ from the background the more salient it is. Matrix \( A \) adds some context properties to the vector \( b \). Consider a matrix \( A \) built of a vector \( \hat{n} = (n_1, n_2, \ldots, n_N) \)

\[
A = \begin{pmatrix}
1 - n_1^2 & -n_1 n_2 & \cdots & -n_1 n_N \\
-n_2 n_1 & 1 - n_2^2 & \cdots & -n_2 n_N \\
\vdots & \vdots & \ddots & \vdots \\
-n_N n_1 & -n_N n_2 & \cdots & 1 - n_N^2
\end{pmatrix}
\]
We can observe that if one element in \( \hat{n} \) is dominant, e.g., \( n_1 = 1 \), then:

\[
A = \begin{pmatrix}
0 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{pmatrix}
\]  

(6)

If all elements are equal, \( \hat{n} = \left( \frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}} \right) \), then:

\[
A = \begin{pmatrix}
1 - \frac{1}{N} & -\frac{1}{N} & \cdots & -\frac{1}{N} \\
-\frac{1}{N} & 1 - \frac{1}{N} & \cdots & -\frac{1}{N} \\
\vdots & \vdots & \ddots & \vdots \\
-\frac{1}{N} & -\frac{1}{N} & \cdots & 1 - \frac{1}{N}
\end{pmatrix}
\]  

(7)

From equation (6) we can derive that \( b_1 \) will be uninteresting if \( n_1 = 1 \). From equation (7) we can derive that only \( M << N \) elements in \( \hat{n} \) can be dominant if no element in \( \hat{n} \) is dominant.

In other words:

- If the background is e.g. red then a red foreground area cannot be salient.
- A large feature dimension (\( N >> 1 \)), e.g., a texture dimension, can have strong responses on several textures at the same time at one area.
- A small feature dimension on the other hand, e.g. an intensity dimension, can only have strong response on one feature at one time and area (an area is dark or bright).

Hence, the matrix \( A \) adds some contextual properties to the vector \( b \).

5 The Spotlight Mechanism

After having propagated the feature values up the pyramid, and calculated saliency values, each node has a saliency value \( -1 \leq S \leq 1 \). In the top layer we select the most salient node and form a spotlight. The spotlight is formed by selecting all nodes with a value greater than zero and that is connected to a selected node in the layer above, see image 3.

In the bottom layer a set of pixels will be selected, that is the region of interest. The receptive field of the top node is normally larger than the selected region of interest. All of the pixels in the receptive field are inhibited to prevent the system from returning attention to the same area.

6 Results

In the implementation for the experiments a pyramid with two layers is used. The input feature images have the size 720*576 pixels. The nodes in the bottom layer are connected to an area of 5*5 pixels in the feature images. The nodes are separated by 5 pixels, so there is no overlap. The top layer has one of two configurations:
Fig. 3. The spotlight mechanism

1. The wide area scan: The top layer nodes are connected to an area of 60*60 nodes (a receptive field of 300*300 pixels, about 1/4 of the image size).
2. The narrow area scan: The top layer nodes are connected to an area of 30*15 nodes (a receptive field of 150*75 pixels, about the size of the items on the table).

For the experiments we used a picture of our colleague making a toast (figure 4 top left).

We want to verify how the solution integrates cues and handles scales, so we applied three different tasks:

1. We let all features have equal weight in a wide area scan. The left side of the white wall plus our colleagues right arm is much brighter than the surrounding, and thus captures the attention (see figure 4 top right).
2. We search for a wide white and blue area and found the white wall and the blue sofa. In the first run we found the left side and in the second we found the right side. The result is similar to the previous search (where no specific features was searched), except that the sofa is included. In the market the background and the sofa will "sell" their features mainly to the red shirt and dark shadows (and green table in the 2nd search) below the attention area (see figure 4 middle left and right).
3. We searched for a narrow white and blue area and found the two plates in the top-left corner of the table. In the next run we found the plate in the lower-left corner of the table. We note that the white and blue plates will "sell" their features mainly to the dark surrounding to the left of the table and the green table, (see figure 4 bottom left and right).
Fig. 4. (top left) Input image. The background is white, the sofa is blue, our colleague has a red shirt, the table is green, and the plates to the left of the table are white and blue. (top right) Search for any salient wide area (middle) 1.st and 2.nd search for a wide white and blue area, (bottom) 1.st and 2.nd search for a narrow white and blue area.
7 Summary

A framework for visual attention has been discussed. The framework calculates saliency, with respect to a given task, using a multi-scale pyramid and multiple cues. The saliency computations is based on game theory concepts, or specifically a coalitional game. The most salient area is the region of interest.

The experiments basically reports expected results. However it should be noted that the task specification has been statically set by the authors from empirical study, important improvements can be achieved by context aware tasks and the potential to improve tasks by learning.

Acknowledgement

This research has been sponsored by the EU through the project "Cognitive Vision Systems" – IST-2000-29375 Cogvis. This support is gratefully acknowledged.

References