A Novel Environment Perception Algorithm Combining Biologically Inspired Features and Brain-like Memory Pattern

Xingwei Yan¹, Afeng Yang¹, Chun Du² and Zhangmeng Liu¹

¹ State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology, No.109, Deya Road, Kaifu District, Changsha, China.
² College of Electronic Science, National University of Defense Technology, No.109, Deya Road, Kaifu District, Changsha, China.
Email: yanxingwei@nudt.edu.cn; yangafeng09@163.com, dc.dd@163.com

Abstract. In this paper, we proposed a novel framework for environment perception. The proposed model combines the physiological mechanisms of human vision and psychological principle of cognitive development. Our algorithm consists of two major components, a gist module and a memory module. The gist module simulates the cognition process associated with visual attention. The memory module is simulated by the Incremental Hierarchical Discriminant Regression (IHDR) tree, which ensures the memories of the world are distributed over the hierarchy. Experimental results on the USC dataset demonstrated that the proposed algorithm is robust and efficient.

1. Introduction
Online environment perception is vital for many robots’ applications such as simultaneous localization and mapping (SLAM) and visual navigation[1-2]. Traditional methods tried to solve the problem with a multi-sensor configuration for indoor environment. The strategy has been demonstrated to be effective as indoor environment has salient spatial and structure regularities. However, these algorithms often fail the outdoors, where the scene environment is more complex[3-4]. Inspired by the recent progress from the computer vision community, more and more researchers start to tackle the problem from a different perspective by mapping environment perception as scene recognition using visual sensors[5-7]. Remarkable progress has been made to improve the speed and robustness of scene recognition in the last few years. However, the fundamental problems for reliable scene recognition are still far from being completely solved. Scene recognition from moving platforms is challenging task because of a wide range of possible viewpoints and illumination conditions, dynamic background, moving cameras and the other independently moving objects in the field of view.

Early works on scene recognition use low-level features directly from the whole image or from a fixed spatial layout, combining with supervised learning methods to classify images into several semantic classes. Later work focusing on the semantic modelling of scene recognition can be primarily categorized into bag-of-words models and object-based methods. More recently, re-searchers are paying more and more attention to computational models, which are derived from the psycho-logical and physiological mechanisms of the human vision[8-12]. Several biologically inspired approaches based on the gist model have been developed[11-12]. The gist model[13] uses simple visual features shared with attention system in an overall biologically-plausible framework. Autonomous mental development theory has been used for environment interactions description as well[14-15]. Most recently, researchers have demonstrated that a combination of computational models from both human
neuroscience and cognitive development will certainly improve the overall scene recognition performance [16-18].

In this paper, we want to improve the scene recognition performance by utilizing multiple cues, while placing emphasis on the learning and development of human brain. We propose a novel framework for scene classification, which combines biologically inspired features and brain-like memory pattern. Our algorithm consists of two major components, a gist module and a memory module. The gist module (extracting intensity features, color features and orientation features) simulates the cognition process associated with visual attention. The color and luminance features are computed by a set of center-surround operations akin to visual receptive fields, and the orientation features are represented as C1 units [13], which correspond to complex cells in the visual cortex. The memory module is simulated by the Incremental Hierarchical Discriminant Regression (IHDR) tree [18], which ensures the memories of the world are distributed over the hierarchy. On one hand, a cognitive model of the world is constructed through autonomous real-time interactions with environments. On the other hand, scene recognition is performed by recalling reasoning of memory.

The gist module summarizes the quintessential characteristics of scene image, and uses the low dimensional biologically-inspired features to describe the scene. The memory module is a coarse approximation of computational model for the development of the memory. It is to establish the association between the gist feature and scene label simulating the way of how our human memory works. The proposed framework absorbs the best knowledge of both physiology and cognitive psychology, and is very biologically motivated. In addition, the number of samples needed at the training stage is very small.

The rest of this paper is organized as follows. Section 2 gives a detailed description of the gist module for effective scene representation. Section 3 presents a general analysis on the memory module and implements an IHDR tree to perform discriminant analysis incrementally. Experimental results are provided in Section 4. Finally, Section 5 concludes the paper.

2. Gist Module

Human can summarize the quintessential characteristics of an image just with a glance. It is an exquisite ability capturing the “gist” of a scene in less than 100ms [19]. Gist computations occur in brain regions, which respond to “places”. It prefers scenes, which have salient spatial layout, and it is also affected by spectral contents and color diagnosticity. The gist model used in our frame-work takes advantage of the research findings by [11] and [13], which unifies the biologically inspired feature such as C1 units, color and intensity units together to represent scene image in an inexpensive manner. There is only one visual cortex in the primate brain, which serves both saliency and gist computations. Siagian and Itti [11] proposed a low-level biologically-inspired feature extraction mechanism, which coarsely corresponding to cortical visual areas V1 through V4 and MT. Visual input is first decomposed into a set of distinct feature channels using linear filters. The number and response of the linear filters have been chosen according to their neuronal equivalents in the early stages of visual processing in primates. The decomposition has also been performed at different spatial scales to represent objects of different sizes in separate subdivisions of the channels.

2.1 Luminance Features

The first set of feature maps is concerned with luminance contrast, which, in human, is detected by neurons sensitive either to dark centers on bright surrounds or to bright centers on dark surrounds. Six different pairs of center and surround spatial scales are used to compute the luminance feature maps \( I(c, s) \)

\[
I(c, s) = \left| I(c) - I(s) \right| \tag{1}
\]

The second set of maps is constructed for the color channels, which are represented using color double-opponent system. In the center of their receptive fields, neurons are excited by one color (e.g., green) and inhibited by another (e.g., red), while the converse is true in the surround.

\[
RG(c, s) = \left| (R(c) - G(c)) - (R(s) - G(s)) \right| \tag{2}
\]
\[ BY(c,s) = (B(c) - Y(c))\Theta(B(s) - Y(s)) \]  

### 2.2 Orientation Features

Local orientation information is obtained from gray-value image using Gabor filters, approximating the receptive fields sensitivity profile of orientation-selective simple cells in S1 layer of primary visual cortex \cite{20}. We arrange the S1 filters to form a pyramid of eight scales and span a range of sizes from 7-by-7 to 21-by-21 pixel. Four orientations (0°, 45°, 90°, 135°) are considered to represent different orientation-selective features. In the standard model of visual cortex \cite{20}, simple S1 units alternate with complex C1 units. C1 units pool over S1 units and only keep the max response of a local area of S1 units from the same orientation and two adjacent scales to increase invariance. The size of local area is decided by the scale band index of S1 units. We arrange 4 bands from 8-by-8 to 14-by-14 grid with the step of two pixels, thus 16 feature maps of C1 units are obtained. The C1 features are robust because that maximum operation highlights the orientation information.

### 2.3 Gist Feature Representation

After the visual cortex features are computed, each sub-channel extracts a gist vector from its corresponding feature map. There are 12 color feature maps, 6 luminance feature maps, 16 orientation feature maps together to represent an image. And averaging operations, which can be seen as the simplest neural plausible computation, is applied in each sub-region. Finally, 16 mean values are utilized to represent each feature map and 544 values are obtained for gist representation. The gist extraction method based on first-order statistics is proved to be sufficient to yield reliable classification \cite{11} for its stability in averaging out local and random noise.

### 3. Memory Module

The IHDR tree \cite{18} incrementally builds a decision tree for very high dimensional decision spaces by an online real-time learning system. It can simulate the hierarchy of the cortex, which ensures the memories of the world are distributed over the hierarchy. To state the mechanism of IHDR, we first define the task of scene classification as follows:

Given a training sample set \[ L = \{(x_i, l_i) | i = 1, 2, \ldots, n\} \], where \( x_i \in X \) is an input feature vector and \( l_i \) is the scene symbolic label of \( x_i \), the classification task is to determine the class label of any unknown input \( x \in X \). IHDR casts a classification task into a regression one by using class means from input space as output labels. Every sample \( (x_i, l_i) \) that belongs to class \( l_i \) is converted to \( y_i \), which is the mean of all \( x_i \) belongs to the same class. The mean is updated by using amnesic average to perform the learning in an incremental way. The amnesic average \cite{39, 40} is motivated by the scheduling of neuronal plasticity, which adaptively changes with on-going experience. The new input gets more weight than old inputs as given in the following expression:

\[ \frac{x^{(t)}}{x} = \frac{t - 1 - \mu}{t} x^{(t-1)} + \frac{1 + \mu}{t} x_t \]  

Where \( x^{(t-1)} \) is the previous average, \( x_t \) is the new input arriving at the time \( t \), \( \mu \geq 0 \) is an amnesic parameter.

The IHDR tree can mimic the systematic organization of long-term memory and short-term memory. The shallow nodes store the primitive prototypes, which are not visited often, corresponding to the long-term memory that forget and neglect unrelated details for memory efficiency and generalization. The micro-clusters in leaf nodes are visited often, corresponding to the short-term memory that the details are forgotten through incremental averaging. The structure of IHDR tree can be seen in Fig.1, which is a hierarchical tree that has three kinds of node: root, internal node and leaf node.
Two types of clusters are incrementally updated at each node: $y$-clusters, which are clusters in the output space $Y$, determining the virtual class label of each training sample $(x, y)$ based on its $y$ part; $x$-clusters, which are clusters in the input space $X$, approximating the sample population in $X$ space. For each internal node, there are a maximum of $q$ clusters for each type. The $q$ centers which generate $q-1$ discriminating features spanning a $(q-1)$-dimensional space. The statistics of each $x$-cluster is updated by the amnesic average.

At the training stage, the samples are firstly clustered into root according to its virtual class label of $y$ part, and $n$ subclasses are formed. The training samples are then clustered into $n$ subclasses based on its $x$ part, and $n$ internal nodes are spawned, which are child nodes from the root. If a close approximation is required, it may spawn children nodes from the current node. When the number of samples in a node is too small to give a good estimate of the statistics of $q$ $x$-clusters, the node is a leaf node.

At the testing stage the root of the tree and sample $x$ are given, we compute the sample size dependent negative-log-likelihood (SDNLL) distance \cite{17} between sample $x$ and children nodes and finds the best matching children node. The iteration stops until getting the leaf node and the output label is assigned to be sample $x$.

4. Experimental Results

The performance of the proposed algorithm has been evaluated using the University of Southern California (USC) scene dataset \cite{11}, which contains video clips of three sites on campus. The three sites are Ahmanson Center for Biological Research (ACB), Associate and Founders Park (AnF), Frederick D. Fagg park (FDF). The video clips are divided into segments for classification and each site scene consists of 9 segments. Each segment represents a portion of a hallway, path, or road interrupted by a crossing or a physical barrier at both ends.

4.1 ACB Dataset

Firstly, we want to test our algorithm in a rigid and less spacious man-made environment. Each segment is a straight line and part of a hallway. Some segments are shown in Fig.2.
We combined all test video clips of four lighting conditions together to evaluate the performance of the proposed method. We uniformly sampled 10 images per second from video clips. The total sample dataset consists of 14158 frames. To test the effectiveness and robustness of the scene classification system, the sample dataset has been randomly divided into the training sample set and the testing sample set. The proportion of training set to testing set has been set as 1/2, 1/8, and 1/25 in three independent experiments. Table 1 shows the statistics of the experimental results.

| Segment Number | Experimental Results (accuracy) |
|----------------|---------------------------------|
|                | 1/2    | 1/8    | 1/25  |
| 1              | 0.9992 | 0.9998 | 0.9995|
| 2              | 1.0000 | 0.9987 | 0.9976|
| 3              | 1.0000 | 0.9994 | 0.9989|
| 4              | 0.9977 | 0.9983 | 0.9977|
| 5              | 0.9988 | 0.9991 | 0.9905|
| 6              | 1.0000 | 0.9977 | 1.0000|
| 7              | 0.9982 | 0.9980 | 0.9912|
| 8              | 0.9975 | 0.9991 | 0.9939|
| 9              | 1.0000 | 1.0000 | 1.0000|
| Total(1-9)     | 0.9991 | 0.9987 | 0.9968|

It can be seen that the proposed algorithm performs very well with different experimental setups.

4.2 AnF Dataset
Secondly, we tested the proposed algorithm in a vegetation-dominated environment. Some segments from the AnF dataset are shown in Fig.3.

![Figure 3. Segments images from AnF.](image)

We use the same experimental setting as in the ACB dataset. The total sample dataset consists of 30094 frames. Table 2 shows the statistics of the experimental results.
Table 2. Experimental results on AnF dataset.

| Segment Number | Experimental Results (accuracy) | 1/2   | 1/8   | 1/25  |
|----------------|---------------------------------|-------|-------|-------|
|                |                                 | 0.9946| 0.9912| 0.9833|
| 1              |                                 | 0.9964| 0.9909| 0.9936|
| 2              |                                 | 0.9978| 0.9945| 0.9949|
| 3              |                                 | 0.9959| 0.9938| 0.9935|
| 4              |                                 | 0.9960| 0.9919| 0.9919|
| 5              |                                 | 0.9985| 0.9954| 0.9925|
| 6              |                                 | 0.9973| 0.9893| 0.9911|
| 7              |                                 | 0.9978| 0.9956| 0.9966|
| 8              |                                 | 0.9981| 0.9953| 0.9974|
| 9              |                                 | 0.9971| 0.9935| 0.9928|
| Total(1-9)     |                                 | 0.9971| 0.9935| 0.9928|

In [11] and [13], the classification results on this experiment are much lower than the other two datasets as the environment is less structured and heavy shadowed. Compared with the results on ACB dataset, the performance of our algorithm deteriorates slightly as the AnF dataset is more challenging.

4.3 FDF Dataset

Thirdly, we tested the proposed framework in a more dynamic environment. The FDF dataset consists of scenarios where dynamic objects such as people and cars exist. And there are also partial and severe occlusion in the scene images. Some segments from the FDF dataset are shown in Fig.4.

![Figure 4. Segments images from FDF](image)

All test video clips of four lighting conditions are combined together to test the performance of the proposed method, there are 43984 sample frames in total. We use the same experimental setting as the other two datasets. Table 3 shows the statistics of the experimental results.
Table 3. Experimental results on FDF dataset.

| Segment Number | Experimental Results (accuracy) |
|----------------|---------------------------------|
|                | 1/2    | 1/8    | 1/25   |
| 1              | 0.9993 | 0.9933 | 0.9928 |
| 2              | 0.9971 | 0.9968 | 0.9921 |
| 3              | 0.9952 | 0.9934 | 0.9913 |
| 4              | 0.9975 | 0.9981 | 0.9941 |
| 5              | 0.9977 | 0.9986 | 0.9925 |
| 6              | 0.9991 | 0.9964 | 0.9961 |
| 7              | 0.9973 | 0.9945 | 0.9947 |
| 8              | 0.9977 | 0.9944 | 0.9896 |
| 9              | 0.9970 | 0.9949 | 0.9943 |
| Total (1-9)    | 0.9977 | 0.9951 | 0.9932 |

5. Conclusion
Inspired by the recent advance in human cortex, we proposed a novel framework for scene recognition. The proposed model combines the physiological mechanisms of human vision and psychological principle of cognitive development. Our algorithm consists of two major components, a gist module and a memory module. The gist module simulates the cognition process associated with visual attention. The memory module is simulated by the IHDR tree, which ensures the memories of the world are distributed over the hierarchy. On one hand, a cognitive model of the world is constructed through autonomous real-time interactions with environments. On the other hand, scene recognition is performed by recalling reasoning of memory. Experimental results on the USC dataset demonstrate the robustness and effectiveness of the proposed method.

6. Acknowledgement
This work was supported by the National Natural Science Foundation of China under Grant No.61601481.

7. References
[1] Leonard J J and Durrant-Whyte H F 1991 *IEEE Transactions on Robotics and Automation* 7 376–382
[2] Thrun S, Burgard W and Fox D 1998 *Machine Learning* 31 29-53
[3] Blaer P and Allen P K 2002 Topological mobile robot localization using fast vision techniques *IEEE International Conference on Robotics and Automation, 2002. Proceedings. ICRA* pp 1031-1036 vol.1
[4] Torralba, Murphy, Freeman and Rubin 2003 Context-based vision system for place and object recognition *Proceedings Ninth IEEE International Conference on Computer Vision* pp 273-280 vol.1
[5] Serrano N, Savakis A E and Luo J 2004 *Pattern Recognition* 37 1773-1784.
[6] Vailaya A, Figueiredo M A, Jain A K and Zhang H J 2001 *IEEE Transactions on Image Processing* 10 117-130
[7] Chang E Y, Goh K, Sychay G and Wu G 2003 *IEEE Transactions on Circuits and Systems for Video Technology* 13 26-38.
[8] Itti L, Koch C and Niebur E 1998 *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 1254-1259
[9] Itti L and Koch C 2001 *Nature Reviews Neuroscience* 2 194-203
[10] Begum M and Karray F 2011 *IEEE Transactions on Autonomous Mental Development* 3 92-105
[11] Siagian C and Itti L 2007 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29** 300-312

[12] Siagian C and Itti L 2007 Biologically-inspired robotics vision monte-carlo localization in the outdoor environment *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems* pp 1723-1730

[13] Song D and Tao D 2008 C1 units for scene classification *2008 19th International Conference on Pattern Recognition* pp 1-4

[14] Tomaso P and Emilio B 2004 *Nature* **431** 768-774

[15] Weng J, Mcclelland J, Pentland A, Sporns O, Stockman I, Sur M and Thelen E 2001 *Science* **291** 599-600

[16] Weng J 2012 *IEEE Transactions on Autonomous Mental Development* **4** 29-53

[17] Weng J and Hwang W S 2003 *International Journal on Document Analysis and Recognition* **5** 118-125

[18] Juyang W and Wey-Shiuan H 2007 *IEEE Transactions on Neural Networks* **18** 397-415

[19] Oliva A and Schyns P G 1997 *Cognitive Psychology* **34** 72-107

[20] Serre T, Wolf L, Bileschi S, Riesenhuber M and Poggio T 2007 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29** 411-426