Abstract – An important component of higher level fusion and decision making is knowledge discovery. One form of knowledge representation is a set of probabilistic relationships between entities. Here we present biologically-inspired algorithmic support for automatic scene understanding and complex object recognition. Our algorithm learns the association between scene and complex objects and their primitive components with and/or without a priori knowledge. In addition, the spatial relationships between the simple constituents and their probabilities are learned incrementally. Complex Object Associative Learning Enables SCene Exploitation neural network (COALESCE) is a hybrid neural network based on probabilistic associative learning and hyper-elliptical learning algorithms. The Object Probabilistic Associative Learning (OPAL) algorithm automatically discovers the conditional probabilities and hierarchical structure comprising a scene. Hyper-Elliptical Learning and Matching (HELM) learns spatial relationships between objects in an object-centric reference frame.

Keywords: Scene understanding, complex object recognition, learning, detection, neural networks.

1 Introduction

Attempts to achieve higher level fusion and decision-making have been made for scene understanding and complex object recognition using a variety of techniques, including rules-based reasoning, logic-based methods, Bayesian networks, fuzzy logic, and neural networks [1]. We propose a new approach to the problems of knowledge representation and learning for scene understanding. This approach combines several elements from the field of neural network modeling including associative learning [3] based on Hebbian neural networks [6], and hyper-elliptical category learning algorithm [5]. Although alternative methods have dealt with scene understanding and complex object recognition scenarios, the goal here is to demonstrate the potential of our approach rather than to evaluate comparative performance between approaches. Successful demonstration, as presented here, establishes the foundation for future development to tackle more difficult scenarios beyond scene understanding, such as recognition of objects with common features.

2 Related work

Many vision scientists have argued that in order to perform object recognition it is necessary to learn representations of the underlying components of images [1]. Recognition by Components (RBC) theory has hypothesized that objects are classified and then recognized by spatial arrangement of their primitive components [1]. Such primitive components correspond to objects, object-parts, or features. The RBC theory was extended to compound object structure of a scene [2], as one can imagine that a scene contains a number of different objects, which occur in different configurations to form a distinct image. For example, a playground may be identified by its constituents: slides, swings, sandboxes, merry-go-round, etc. However, unlike simple objects, compound objects do not have simple and well-defined geometric models to describe them. Instead, the important characteristics of a scene are the dependencies among objects which tend to co-occur, and their spatial relations which are inherently more variable than for primitive components of an individual object.

Knowledge-based vision systems, such as Hidden Markov Models (HMM) and Bayesian Networks (BN), have been shown to be effective for scene understanding and compound object recognition. These models provide a reusable and extensible semantic ontology as well as a great capacity for image interpretation and detail examination. However, the major negative point of these systems is their reliance on a priori knowledge and conditions – these are difficult and time consuming to produce. Bayesian networks have often been applied for object recognition and scene understanding, using the evaluation probabilities and confidence levels of each constituent [8]. However, dependencies among objects

---

1 The material in this paper is based upon work supported by the AFOSR under Contract No. FA9550-06-C-0018.
are often context-dependent and difficult to access, undermining the validity of such networks.

The approach presented here proposes a novel approach to recognize a scene without any handcrafted \textit{a priori} knowledge. Complex Object Associative Learning Enables SCene Exploitation (COALESCE) is a biologically-inspired algorithm for automatic scene understanding and complex object recognition. It incrementally learns, with and/or without \textit{a priori} knowledge, the associations and interdependencies between compound objects and their primitive components. In addition, the spatial relationships among the simple constituents and their probabilities of pair-wise occurrences are learned incrementally.

### 3 Proposed approach

COALESCE is a hybrid neural network based on probabilistic associative learning and hyper-elliptical category learning algorithms. The first component of COALESCE, Object Probabilistic Associative Learning (OPAL) is based on Probabilistic Neural Associative Incremental Learning (pNAIL) [4] and automatically discovers the conditional probabilities and hierarchical structure of elements comprising a scene. COALESCE’s second component, Hyper-Elliptical Learning and Matching (HELM), learns spatial relationships between constituent objects in an object-centric frame of reference. The two components operate in parallel during learning. During detection OPAL outputs a list of possible scenes according to observed simple objects, which are then evaluated by HELM according to their spatial relationships.

Figure 1 shows the block diagram architecture of COALESCE. OPAL receives binary data that reflects the existence (1) or nonexistence (0) of an object in a scene during learning and detection. HELM accepts a combination of analog and binary input corresponding to distance, direction/orientation, above/below, out/in, and/or other desired spatial relationship features. During learning OPAL and HELM learn their mappings in parallel. When in detection mode, OPAL’s output is a list of possible scenes based on the set of simple objects present in the scene. Each item in this list is then evaluated by HELM. If the spatial pattern of objects within a scene matches the previously learned pattern for a scene, positive detection is achieved. The detailed modes of learning and detection operation of OPAL and HELM are described in sections 3.1 and 3.2.
3.1 OPAL

Each node in an OPAL network represents a semantic item (e.g., an object, an entity, a scene, or a complex object) and the output of each neuron is linked to the input of other neurons in the network, as shown in Figure 2. In addition to the nodes representing simple objects (blue nodes) and scenes (green nodes) there are other nodes (black nodes) representing combinations of simple objects that are active only when all their simple object constituents are active (green arrows). There are also inhibitory objects or entities (red node) that inhibit the scene object when active. The links between nodes are represented as weighted connections between pairs of nodes. The approach to learning links between objects and predicting the resulting scene is based on the associative learning algorithm introduced in [3]. The connection weights between objects change via gated Hebbian learning, which are updated according to a modified outstar learning law [6]:

$$\Delta w_{ik} = \frac{1}{N_{jk}} \cdot x_{ik} \cdot (x_{ik} - w_{ik})$$

where $N_{jk}$ is the number of times that node $j$ has been activated in the $k$th set of weights (that corresponds to the simple objects indexed by $k$), $w_{jk}$ is the connection weight from node $j$ to node $i$, and $x_{ik}$ and $x_{ik}$ are the activations of simple object $j$ and scene $i$ respectively. The learning rate $(1/N_{jk})$ is node-dependent, such that it decreases with the amount of activity that has been encountered by a node. For a node $j$ in the $k$th set of weights, the learning rate first starts at a maximum of 1 and then decreases inversely with $N_{jk}$. Each node thus begins in a fast-learning mode, and then the weights are slowly tuned as more data is presented. Learning is pre-synaptically-gated by activation at the source location. If this location is not active, then no connections from this location to other locations will change their weights. If the source location is active, then links with any active target locations will increase their weights and links with any inactive target locations will decrease their weights. Given the binary activations used in the network, weights are bounded between 0 and 1 and the size of weight changes is governed by the learning rate and the size of the current weight. We have shown empirically (unpublished results) that this data-dependent learning rate causes the learned weights to accurately track the conditional probabilities encountered in the training data. Future work will present these results and seek to derive this property analytically.

This form of learning has a number of attractive properties for the current application. First, more frequent combinations of simple and scene objects are rapidly learned, as indicated by larger weights. Second, random/infrequent combinations will cause learning when they occur, but will also be unlearned through weight decay when they do not occur. This property provides a form of noise tolerance. Third, the system is also able to maintain multiple sets of models for different contexts.

For each observation of a set of simple and scene objects, pNAIL can make predictions about possible scenes with weights indicating probabilities of simple objects (and their combinatorial) having previously been observed in a scene. Of course, there may be many non-zero weights emanating from the simple object nodes, a large fraction of which may be small (having arisen from only a few observations). Given the probabilistic nature of the learned weights, a minimum probability threshold may be set to require a minimum level likelihood in order to make a prediction.

![Figure 2: OPAL Architecture for detecting a complex entity (CE) and its relation to four simple objects (a, b, c, d) and their combinatorial objects (ab, ac, ...). Presence of node IE inhibits the activation of CE, and hence eliminating it from the list of possible outcomes.](image)

Table 1: Calculated maximum number of nodes in an OPAL architecture as a function of simple objects

<table>
<thead>
<tr>
<th># of Simple Objects ($n$)</th>
<th>Total # of Nodes ($M$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>1023</td>
</tr>
<tr>
<td>15</td>
<td>32767</td>
</tr>
<tr>
<td>20</td>
<td>1048575</td>
</tr>
</tbody>
</table>

The total number of nodes, including simple nodes and their combinations, is governed by:

$$M = \sum_{r=1}^{n} \frac{n!}{r!(n-r)!}$$
only in a small number of simple objects; therefore the total number of nodes in use will not exceed more than several hundred (when \( n < 10 \) simple objects). Therefore, in practice OPAL is a sparsely connected network of nodes operating in parallel.

### 3.2 HELM

To learn context-sensitive models of spatial relationships among simple objects observed in a scene, we have developed a neural network classifier which incrementally constructs a multidimensional Gaussian (hyper-ellipsoid) model of each object’s spatial map of spatial relationships to all other objects in a scene. The Hyper-Elliptical Learning and Matching algorithm (HELM) learns spatial relationships between objects in an object-centric frame of reference that is relatively insensitive to outliers. Figure 3 shows an object-centric frame of reference for the simple object 1 located at the origin marked by \( X \), and its relative spatial map to objects 2-5 (ellipses) including distances and their relative proximity (e.g., above/below, out-in and relative direction). In this representation, relative object locations may comprise a point, a line, or a (multi-dimensional) area. For example, in Figure 3 object O4 (blue ellipses) has been observed to be outside of object O1 several times.

HELM can learn the normal feature pattern (spatial proximity, distance and relative direction) along with other feature dimensions comprising the learning hyperspace. When a new data point (observation of an object in a scene) is acquired, the network updates its parameters adaptively to the incoming data, and provides an accurate measure of normal distance and proximity. During detection, the network checks if the reported spatial relationships between the data point belonging to an object and the object at the origin are within the predefined (controllable) threshold of the hyper-ellipse categorizing that object (straight line to the center of an ellipse in Figure 4). This difference is measured by the Mahalanobis distance in units of standard deviations \([7]\).

If a point falls outside of the object’s learned category (hyper-ellipse), then that object is not considered as part of the hypothesized scene and does not provide evidentiary support for that scene. Note that ellipses represent the statistical representation of data learned incrementally. The speed and performance of this learning algorithm makes it suitable for real-time situations wherein an operator/analyst can interactively facilitate the learning process and/or exercise control over the sensitivity level of the system to control false alarms.

### 3.3 Learning

Learning of a scene is driven by presence of constituent simple objects and of the distribution of spatial relationships between constituents, which can be acquired from a ground-truth dataset. Both OPAL and HELM representations are learned. Learning is supervised and rapid; it can be achieved via a limited number of examples\(^2\). OPAL learns the bottom-up and top-down weights which represent probabilities of co-occurrence between pairs of objects, i.e. between simple objects (as well as their combinatorial objects) and scene objects (Figure 2). HELM learns spatial relationships between

---

\(^2\) The algorithm can learn with having only one exemplar, however in order to have a meaningful weight that correspond to conditional probabilities more exemplar are needed. In this example, five examples were used.
3.4 Detection

For an unknown scene, the detection results contain the name of the scene and the probability of successful match. Scene understanding takes place in two stages, as shown in Figure 1. First, simple objects are presented to OPAL to narrow scene candidates based on objects present in the data point. Scenes with strongest activations are identified by thresholding association weights. Leading candidate scenes with highest weights (probability of co-occurrence) to observed simple objects are subjected to further examination by HELM.

Second, HELM refines scene predictions based on spatial relationships between objects comprising hypothesized scenes. Each candidate scene consists of several simple object-centric models (Figures 3 and 4). Simple object distance from model mean is measured in Mahalanobis distance (Figure 4) in each object-centric frame of reference. Average distance is compared to a threshold to determine match or mismatch to the model. The degree of complex object match depends on set of object-centric model matches. Perfect match occurs when all object-centric models return a match. When no match is detected, COALESCE may be instructed to learn the scene and its observed simple objects.

3.5 Example

As an example, a playground is chosen to be identified by its constituents: slide, swing, sandbox, merry-go-round, and club-house. However if a building is nearby, then the scene may be a school or a house, i.e., confusing scenes. To test the COALESCE model, data from five exemplars of playground were extracted from the Google-Earth images in Massachusetts, USA. All exemplars of playground had the simple objects listed above; however a truth set was created for scenarios such as ‘What if object X was not observed’ (as depicted in the rows of Table 2). Columns in Table 2 represent the single objects (SL=slide, MG=merry-go-round, SW=swing, SB=sandbox, CH=clubhouse, B=building) and scene objects (SCL=school, PG=playground, H=house).
data without a priori knowledge is an important characteristic of OPAL.

Figure 6 illustrates the six object-centric models of spatial relationship between simple objects (red = slide, green = merry-go-round, black = swing, blue = sandbox, magenta = clubhouse, cyan = building). For clarity only distance and direction features are shown. Each ellipse represents the 95% confidence interval of observed distances and directions. Variability of data along the direction-axis indicates that simple objects are not present in a particular orientation in this scene. Taking the results of OPAL and HELM together, it is deduced that presence of some objects within learned vicinity (e.g., a close by building) in a scene narrows the possible options and enhances correct predictions (e.g., playground).

Figure 6: HELM object-centric models of spatial relationship between the six simple objects in the playground example (red=slide, green=merry-go-round, black=swing, blue=sandbox, magenta=clubhouse, cyan=building). For clarity only distance and direction features are shown. Each ellipse represents the 2 standard deviation about the mean of observed distances and directions.

4 Conclusion

An important component of higher level fusion and decision making is knowledge discovery. One form of knowledge is a set of relationships between entities and their probabilities. COALESCE is a hybrid neural network based on probabilistic associative learning and hyper-elliptical learning algorithms that enables scene understanding and complex object recognition. The OPAL component of COALESCE is a probabilistic associative learning algorithm that automatically discovers the conditional probabilities and hierarchical structure of primitive objects comprising a scene. COALESCE’s second component, HELM, learns spatial relationships between objects in an object-centric frame of reference. The discovery of the knowledge of occurrence and spatial relationships between objects from the truth data without a priori knowledge is an important characteristic of COALESCE. This knowledge discovery from data depends on the appearance of objects a scene captured in the truth data. Results of COALESCE show that presence of some objects within learned vicinity in a scene narrows the possible options and enhances correct predictions.

Our immediate future task is to apply the COALESCE model to a much richer dataset, and compare its results to other commonly used modeling approaches, such as Bayesian networks. In addition, we are planning to apply this algorithm for detecting and classifying complex objects with many common features, e.g. a cup, a pail, and a carrying case all have handles and two have cylindrical shape.

References


