Seminar
Visual Object Recognition

A Study of Design Trade-offs governing Visual Representations
and
A Model of the Structure of Object Memory

Dipl.-Inform. Manfred Hartel
Institute for the History of Science
Physics & Philosophy Group
Overview

1. A Study of Design Trade-offs governing Visual Representations
   ⇒ Minimizing Binding Errors Using Learned Conjunctive Features
      by Bartlett W. Mel & Jozsef Fiser

2. A Model of the Structure of Object Memory
   ⇒ Learning to Detect Objects in Images via a Sparse, Part-Based Representation
      by Shivani Agarwal, Aatif Awan, and Dan Roth
1. A Study of Design Trade-offs governing Visual Representations

1.1 A Method for Analysing the Properties of Feature Sets that support Hallucination-proof Recognition

1.2 Learning Procedures capable of building these Features

1.3 Conclusions
1.1 A Method for Analysing the Properties of Feature Sets that support Hallucination-proof Recognition

1.1.1 The Binding Problem

Hallucination

- Binding is the correct association of detected features to objects

- A binding problem occurs when an object’s representation can be falsely activated by multiple (or inappropriate) other objects.

- Hallucinating is: an object is detected in a scene despite the fact that it is not really present
Combinatorial Explosion

B

too many high-order conjunctions needed

Combinatorial Explosion
detector array

\[ \text{cat}_0 \rightarrow \text{dog} \rightarrow \text{bee} \rightarrow \ldots \]

\[ \vdots \]

\[ \text{cat}_n \]

\[ \text{dog}_n \]

\[ \text{bee}_n \]

\[ \ldots \]

\[ \text{cat}_m \]
1.1.2 Wickelsystems

- Wickelsystems are named after Wickelgren,

- who proposed a scheme for representing the phonological structure of spoken language

- involving units sensitive to contiguous triples of phonological features

- but not to the absolute position of the tuples in the speech stream.
Figure 1: Wickelgren's compromise when an object's representation can be...
Questions:

how recognition performance does depend on:

(1) the number of object categories that must be distinguished,

(2) the similarity structure of the object category distribution (i.e., whether object categories are intrinsically very similar or very different from each other),

(3) the featural complexity of individual objects,

(4) the number and conjunctive order of features included in the representation,

(5) the clutter load (i.e., the amount of visual material in the field of view from which multiple objects must be recognized without explicit segmentation), and

(6) the invariance load (i.e., the set of spatial transformations that do not affect the identities of objects, and that must be ignored by each individual detector).
1.1.3 Methods

1.1.3.1 Text World as a Model for Object Recognition

- a word database

  - containing 44,414 entries, representing all lowercase punctuation-free words and their relative frequencies found in 5 million words in the Wall Street Journal (WSJ) resulting in target objects

  - containing from 1 to 20 parts

  - constructed from a set of relatively few underlying types (26 letters)

  - with highly nonuniform probability distribution ranges from 1 to 400,000
• a **English text database**
  - consisting of approximately 1 million words

• **input text**
  - varying in width from 10 to 50 characters
  - drawn at random from the English text database
1.1.3.2 Recognition Using n-Grams

- A natural class of visual features in text world is the position-invariant n-gram,

- defined here as a binary detector that responds when a specific spatial configuration of n letters is found anywhere (one or more times) in the input array.

- The value of n is termed the conjunctive or binding order of the n-gram,

- and the diameter of the n-gram is the maximum span between characters specified within it.

- For example, th**_ is a 3-gram of diameter 5 since it specifies the relative locations of three characters (including the space character but not wild card characters *), and spans a field five characters in width.
1.1.3.3  Graphical Illustration of Text World Properties

A  Non-Uniform Word Frequencies

B  Non-Uniform Word Sizes

C  Non-Uniform Word Frequencies

D  Non-Uniform Word Sizes
Statistical Dependencies Among Adjacent Letters

- Info Gain About Adjacent Letter (Bits)
- letter to left
- letter to right

|   | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 5 |   |   |   |   |   | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 3 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

Date 26.11.07
File Hartel-Models_Structure_Object_Memory
1.1.4 Results

1.1.4.1 Word-Word Collisions in a Simple 1-Gram Representation

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of 1-grams in R</td>
<td>27</td>
<td>a, b, ... z, _</td>
</tr>
<tr>
<td>Number of distinct words</td>
<td>44,414</td>
<td>From 5 million words in the WSJ</td>
</tr>
<tr>
<td>Word-word comparisons</td>
<td>≈ 2 billion</td>
<td></td>
</tr>
<tr>
<td>Number of ambiguous words</td>
<td>24,488</td>
<td>analogies ≡ gasoline</td>
</tr>
<tr>
<td></td>
<td></td>
<td>suction ≡ continuous</td>
</tr>
<tr>
<td></td>
<td></td>
<td>scientists ≡ nicest</td>
</tr>
<tr>
<td>Largest self-similar cohort</td>
<td>28 words</td>
<td>stare, arrest, tears, restates, reassert, rarest, easter ...</td>
</tr>
<tr>
<td>Baseline entropy in word-frequency distribution</td>
<td>9.76 bits</td>
<td>&lt; log₂ (44,414) = 15.4 bits</td>
</tr>
<tr>
<td>Residual uncertainty about identity of a randomly drawn word, knowing its 1-grams</td>
<td>1.4 bits</td>
<td>Narrows field to ≈ 3 possible words</td>
</tr>
</tbody>
</table>

Table 1: Performance of a Binary 1-Gram Representation in Word-Word Comparisons.
### 1.1.4.2 Word-Word Collisions in an Adjacent 2-Gram Representation

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Adjacent 2-grams in R</td>
<td>729</td>
<td>[aa], [ab], ... [a_], ... [z], [__]</td>
</tr>
<tr>
<td>Number of distinct words</td>
<td>44,414</td>
<td>From 5 million words in the WSJ</td>
</tr>
<tr>
<td>Word-word comparisons</td>
<td>≈ 2 billion</td>
<td></td>
</tr>
<tr>
<td>Number of ambiguous words</td>
<td>57</td>
<td>ohhh, ahhh, shhh, hmmm, whoosh, whirr, …</td>
</tr>
<tr>
<td>Largest self-similar cohort</td>
<td>5 words</td>
<td>ohhh ≡ ohhhh ≡ ohhhhh ≡ ohhhhhh ≡ ohhhhhhhh</td>
</tr>
<tr>
<td>Ambiguous word pairs that are linguistically distinct</td>
<td>4 pairs</td>
<td>asses ≡ assess posses ≡ possesses seamstress ≡ seamstresses intended ≡ indented</td>
</tr>
<tr>
<td>Words with identical adjacent 2-gram multisets</td>
<td>2 words</td>
<td>intended ≡ indented</td>
</tr>
</tbody>
</table>

Table 2: Performance of a Binary Adjacent 2-Gram Representation in Word-Word Comparisons.
1.1.5 An Analytical Model

probability that all of the features feeding a given word detectors are activated by an arbitrary input text:

\[
\begin{align*}
R &: \text{set containing all position-invariant n-grams analysing the input array} \\
d &: \text{total number of feature detectors in } R \\
w &: \text{size of the object’s minimal feature set } \equiv \text{number of features required by the target word detector} \\
c &: \text{number of features activated by a “cluttered” multi-word input text}
\end{align*}
\]
probability that an word \( i \) is hallucinated:

\[
h_i = \frac{(o_i - q_i)}{(1 - q_i)}
\]

\( q_i \): probability that the object does in fact appear in the input text

probability of veridical perception \( \equiv \) that no word \( i \) is hallucinated:

\[
p_v = (1 - h)^N \approx [1 - (c/d)]^N
\]

\( N \): number of words in the word database
dashed lines are predictions of the model
Fits of Analytical Model to Measured Error Rates:
Extended 2-gram Representation

R(2,3) : 1,458 total features : { [aa], [a*a], [ab], [a*b], ...[a_], [a*], ... }

- 2-letter words, r = 2.19
- 5-letter words, r = 2.95
- 7-letter words, r = 3.45

Dashed lines are predictions of the model.


**lessons learned from analysis and experiments :**

1. larger representations can lead to better recognition performance,

2. only those features that are actually used should be included in a representation,

3. considerable statistical redundancy exists in a fully enumerated (or randomly drawn) set of conjunctive features, which should be suppressed if possible,

4. features of low binding order, while potentially adequate to represent isolated objects, are insufficient for veridical perception of scenes containing multiple objects,

5. different objects are likely to have different representational requirements depending on their complexity and

6. the composition of an ideal representation is likely to depend heavily on the particular recognition task.
1.2 Learning Procedures capable of building these Features

1.2.1 Greedy Feature Learning Algorithm

1. Start with a small bootstrap representation. (typically 50 2-grams found to be useful in pilot experiments.)

2. Present a large number of text “images” to the representation.

3. Collect hallucinated words – any word that was ever detected but not actually present – in an input image.

4. For each hallucinated word and the input array that generated it, increment a global frequency table for every n-gram (up to a maximum order of 5 and diameter of 6) that is

   (i) contained in the hallucinated word,

   (ii) not contained in the offending input, and

   (iii) not currently included in the representation.
5. Choose the most frequent n-gram in this table of any order and diameter, and add it to the current representation.

6. Build a connection from the newly added n-gram to any word detector involved in a successful vote for this feature in step 4. Inclusion of this n-gram in these words’ minimal feature sets will eliminate the largest number of word hallucination events encountered in the set of training images.

7. Go to step 2.
1.2.2 Properties of Learned Representations

A Fewer Words are Hallucinated as Representation Grows

![Graph showing the relationship between N-grams included in representation and words hallucinated. The graph compares unsupervised and greedy methods, showing that as the representation grows, fewer words are hallucinated with greedy methods compared to unsupervised methods.]
Fewer Inputs Cause Hallucinations as Representation Grows
1.2.3 Sensitivity of the Learned Representation to Object Category Structure

- In difference to real world objects every word forms its own category.

- Weakening the hallucination criterion to similar rather than identical words.

- E.g. dog $\equiv \{ \text{fog, dig, ...} \}$ but dog $\not\equiv \{ \text{fig, dogs, og, ...} \}$
Decrease in Representational Costs for Broader, Overlapping Object Categories

Number of N-grams

N-gram Order

<table>
<thead>
<tr>
<th></th>
<th>tolerance = 0</th>
<th>tolerance = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1.3 Conclusions

1. Efficient feature sets are likely to

   (1) contain features that span the range from very simple to very complex features,

   (2) contain relatively many simple features and relatively few complex features,

   (3) emphasize features that are only moderately common (giving a representation that is neither sparse nor dense) in response to the conflicting constraints that features should appear frequently in objects but not in backgrounds also composed of objects, and

   (4) in spatial domains, emphasize features that encode the relations of parts to object boundaries.
2. An efficient learning algorithm works to drive toward zero, and therefore to equalize, the false-positive recognition rates for all objects considered individually.

Thus, frequently hallucinated objects – objects with few parts or common internal structure, or both – demand the most attention during learning.
Two consequences of this focus of effort on frequently hallucinated objects are that

(1) the average value of \( w \), the size of the minimal feature set required to activate an object, becomes nearly independent of the number of parts contained in the object, so that simpler objects are relatively more intensively represented than complex objects, and

(2) among objects of the same part complexity, the minimal conjunctive feature sets grow largest for objects containing the most common substructures, though these are not necessarily the most common objects.

A curious implication of the pressure to represent objects heavily that are frequently hallucinated is that the backgrounds in which objects are typically embedded can strongly influence the composition of the optimal feature set for recognition.
3. The demands on a visual representation are heavily dependent on the object category structure imposed by the task.

Where object classes are large and diffuse, the required representation is smaller and weighted to features of lower conjunctive order,

whereas for a category structure like words, in which every object forms its own category that must often be distinguished from a large number of highly similar objects (e.g., cat from cats, cut, rat), the representation must be larger and depend more heavily on features of higher conjunctive order.
2. A Model of the Structure of Object Memory

2.1 Problems in Object Detection

2.2 Image Feature Representation Types

2.3 A Sparse Part-Based Image Representation Approach
2.1 Problems in Object Detection

- the wide range of variations in images of objects belonging to the same object class.

- Different objects belonging to the same category often have large variations in appearance.

- the same object can appear vastly different under different viewing conditions, such as those resulting from changes in lighting, viewpoint, and imaging techniques.

- A successful object detection approach must therefore be able to represent images in a manner that renders them invariant to such intraclass variations,

- but at the same time distinguishes objects of one object class from objects of all other object classes.
2.2 Image Feature Representation Types

- raw pixel intensity representations,
- features obtained via global image transformations,
- local features such as edge fragments,
- rectangle features,
- Gabor filter-based representations,
- wavelet features.
2.3 A Sparse Part-Based Image Representation Approach

2.3.1 Feature Vocabulary

- Extracting interest points from a set of representative images of the target object.

- Interest points are points in an image that have high information content in terms of the local change in signal.

- In this case interest points are:
  - intersections of lines
  - centers of circular patterns

- Information-rich patches are regions surrounding these interest points.

- Similar patches thus obtained are grouped together and treated as a single part.
2.3.2 Image Representation

Representations elements are:

- Parts from the Feature Vocabulary
- Spatial relations among these parts

Spatial relations among these parts:

- distances between pairs of parts
- directions between pairs of parts
AGARWAL ET AL.: LEARNING TO DETECT OBJECTS IN IMAGES VIA A SPAR

Fig. 1. (a) A sample object image used in vocabulary construction. (b) Interest points detected by the Förstner operator. Crosses denote intersection points; circles denote centers of circular patterns. (c) Patches extracted around the interest points.
"cover" new object instances, i.e., to be able to represent new instances using a subset of these patches.

Fig. 2. The 400 patches extracted by the Förstner interest operator from...
PART-BASED REPRESENTATION

Fig. 3. Examples of some of the “part” clusters formed after grouping similar patches together. These form our part vocabulary.

In our experiments, the Förstner operator was applied to a set of 50 representative images of cars, each $100 \times 40$ pixels in size. Fig. 1 shows an example of this process. Patches of
Fig. 4. Examples of the part detection process applied to a positive image (top row) and a negative image (bottom row) during training. Center images show the patches highlighted by the interest operator; notice how this successfully captures the interesting regions in the image. These highlighted interest patches are then matched with
2.3.2.3 Representation of the Feature Vector

\( P_n^{(i)} \) denoting the \( i^{th} \) occurrence of a part of type \( n \) in the image (each \( n \) corresponds to a particular part cluster).

\( R_k^{(j)} (P_m, P_n) \) denoting the \( j^{th} \) occurrence of relation \( R_k \) between a part of type \( m \) and a part of type \( n \) in the image (each \( k \) corresponds to a particular distance-direction combination).
The table shows the occurrence of parts and their corresponding numbers. The notation \( P_n^{(i)} \) indicates a probability or occurrence number, with columns for part numbers and occurrence numbers.

<table>
<thead>
<tr>
<th>Part No</th>
<th>1</th>
<th>i</th>
<th>j</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>q</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
for all values of $j$ this results in an 4-dimensional hypercube
$P_n^{(i)} \rightarrow \mathbb{N}$

$R_k^{(j)}(P_m, P_n) \rightarrow \mathbb{N}$

$1 \quad 27 \quad 138 \quad 543 \quad n$
Thank You for Your Attention
Table of Content

1. A Study of Design Trade-offs governing Visual Representations

1.1 A Method for Analysing the Properties of Feature Sets that support Hallucination-proof Recognition

1.1.1 The Binding Problem

1.1.2 Wickelsystems

1.1.3 Methods

1.1.3.1 Text World as a Model for Object Recognition

1.1.3.2 Recognition Using n-Grams

1.1.3.3 Graphical Illustration of Text World Properties

1.1.4 Results

1.1.4.1 Word-Word Collisions in a Simple 1-Gram Representation

1.1.4.2 Word-Word Collisions in an Adjacent 2-Gram Representation

1.1.5 An Analytical Model

1.2 Learning Procedures capable of building these Features
1.2.1 Greedy Feature Learning Algorithm

1.2.2 Properties of Learned Representations

1.2.3 Sensitivity of the Learned Representation to Object Category Structure

1.3 Conclusions

2. A Model of the Structure of Object Memory

2.1 Problems in Object Detection

2.2 Image Feature Representation Types

2.3 A Sparse Part-Based Image Representation Approach

2.3.1 Feature Vocabulary

2.3.2 Image Representation

2.3.2.1 Elements of the Image Representation

2.3.2.2 Representation of Relations between Parts

2.3.2.3 Representation of the Feature Vector