Visual Categorization: Basic v Super/Sub-ordinate levels

Bhuwan Dhingra
Dept. of Electrical Engg.
Indian Institute of Technology, Kanpur.
email: bhuwand@iitk.ac.in

ABSTRACT

It is widely accepted in the domain of cognitive sciences that there exist hierarchies in categorization – three levels are often posited: super-ordinate, basic and subordinate. Rosch and her colleagues showed in the 1970s that out of these the basic level is accessed first. Recent studies, however, have pointed to evidence that in a rapid visual processing task super-ordinate categories are accessed most rapidly. In this computational study I simulate the hierarchy of categorization using an object recognition classifier. A bag-of-features model is used for feature extraction from object images, followed by k-means clustering to implement categorization. Results show that super-ordinate categories are indeed favoured over basic and subordinate ones. The role of expertise in determining the categories is also studied.

1. Introduction

Since the work of Rosch in the 1970s it has been widely accepted that humans categorize objects at three different levels – *super-ordinate, basic* and *subordinate.* Out of these, Rosch showed through a series of elaborate experiments that the basic level is accessed first in most taxonomies, no matter what perceptual modality is used, [2]. She was able to show in her work on Natural Categories, that given a set of objects (real or abstract) there exists a natural grouping among them which most people conform to. She named these groups as *basic level* categories. Basic level categories can be combined to form *super-ordinate level* categories, or can be divided to form *subordinate level* categories. Hence, while *dog* is a basic level category, *animal* is super-ordinate and *Dalmatian* is subordinate level.

While the idea of these basic categories seems intuitive and natural, there are still a lot of questions unanswered in this area. In particular *how* we extract the basic categories, and *what* these basic categories are is relatively unclear. Rosch considered all sorts of features associated with objects – visual, functional, semantic etc. in her experiments. Since then similar experiments have been conducted with varying perceptual modalities, varying taxonomies and on varying types of subjects to give rise to a vast variety of results.

Some of the results in the study of categorization have contradicted the dominance of the basic level. Recently, there has been an increased interest in the speed and accuracy with which *super-ordinate* categories are accessed. Some studies claim that these are accessed first in a purely visual task, [3,4]. The role of expertise has also been emphasised several times in the formation of these levels, [5]. It is generally expected that expertise in a particular area leads to the dominance of subordinate categories over the basic ones.

In the current study, a computational model working on visual features of objects is developed. The model uses bag-of-features approach to extract features from the object images. The images are then clustered using the k-means algorithm. Three different types of clustering are performed – super-ordinate (*k=2*), basic (*k=4*), subordinate (*k=8*) on the same taxonomy, and the performance for each is computed. Results are studied by varying parameters of the experiment. Finally, the role of expertise is studied by varying the number of images of a particular basic category to see if it splits into subordinate levels.

The rest of the paper is organised as follows – Section 2 provides a brief review of the literature, Section 3 outlines the current model, Section 4 provides the results, and Section 5 provides a discussion of the results along with the conclusions.

2. Previous Work

Rosch conducted her original experiments on Dani children to show the existence of natural categories in [1], and further refined her work to define the three levels in [2]. This drew interest from the entire cognitive community and led to widespread research in the area of categorization. Mervis and Cisafin (1982), [6], showed that children acquire basic level categories first, and adults use the terms corresponding to the basic level categories most often. All these developments led to the hypothesis that basic categories form the mnesic representation inferred first when objects are encountered, and super-ordinate categories are abstract generalizations of these (cf [7]).

Recent studies, however, have challenged this dominance of the basic level categories, claiming that the speed of categorization is much faster for super-ordinate level categories. Marc, Joebert et al performed a rapid visual go/no-go categorization task to study the response time of people when classifying super-ordinate (animal v no-animal) categories compared to subordinate categories (dog v no-dog/bird v no-bird), [7]. Their findings show a marked decrease in response time for the animal v no-animal task. They further suggest that to make a decision at the super-ordinate level only coarse representations are needed, with subsequent refinements for the basic and subordinate levels needing extra processing. In a visual processing task, the coarse features are accessed first and to access the lower levels further information needs to be processed. Most importantly, they emphasise the role of lexical and semantic features in accessing basic categories.

It is intuitive to expect that this hierarchy of categorizations would vary significantly from experts to non-experts in particular domains. Indeed, it has been shown that bird and dog experts can classify the subordinate levels of these categories as quickly as the categories themselves (cf [7]). These studies ultimately suggest that basic level categories are in fact determined by the level of expertise of the subject, and since only a fraction of the population is expert in any given domain there is huge conformity across subjects for these categories.

Several computational models for visual categorization of objects exist in the literature. They differ from one another mainly in the technique used to extract features from the images. Out of these, the most popular technique for extracting features is the Bag-Of-Features (BOF) model. Jiang et al provide a review of the BOF technique and show that the Scale Invariant Feature Transform (SIFT) maximizes the performance for object recognition, [8]. This is the approach that I would be using in the current paper.

3. The Model

A. Bag-of-Features

Rosch claimed the dominance of basic levels was due to the concept of *cue validity*. The cue-validity of a category is defined as the frequency with which an attribute is associated with the category, summed over all attributes of the category. Basic level categories maximize this cue validity, whereas super-ordinate categories have lower cue validity due to absence of common features and subordinate categories have lower cue validity due to lower frequencies of the attributes, [2].

The bag-of feature model for feature extraction from images works on a similar concept. It is a two-stage process – the first stage extracts what are known as *interest points* or *key-points* from the image, and the second stage forms histograms over these extracted points for each image using a *key-point* descriptor. Hence, we can view these *key-points* in the visual domain as the *cues* which Rosch claims define the different levels of categorizations.

In the current model I use the Scale Invariant Feature Transform (SIFT) which bundles together the key-point detector and descriptor. The descriptor is a 128-dimensional vector which describes the spatial structure and orientation distribution around the detected key-point (see [9] for details). Once we obtain these sets of 128-dimensional vectors for each image, they are clustered to form a vocabulary of key-points. Subsequently, images are described by a histogram of the frequency of occurrence of the key-points in the vocabulary in the images. A nearest neighbour technique is used to bin all the key-points of a particular image to the key-points in the vocabulary.

The various functions for implementing the bag-of-features model were obtained from the VLFeat library developed by Vedaldi and Felkurson, [10]. Figures 1a and 1b show two example images with the extracted key-points.

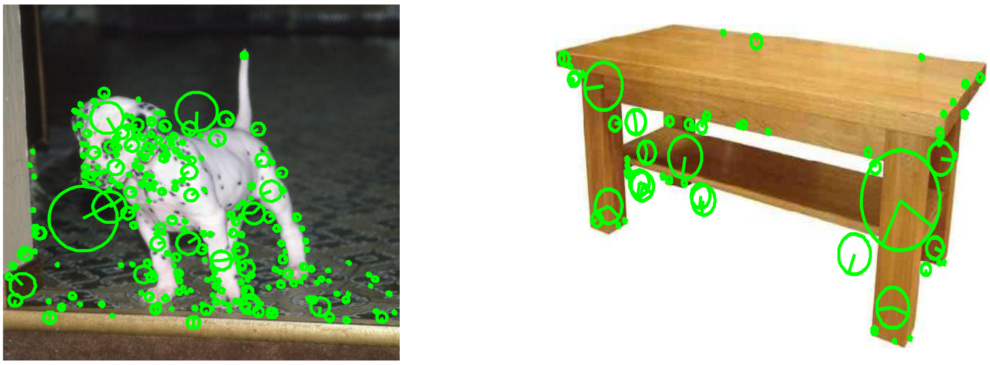


Figure 1a and 1b, Extracted key-points from the SIFT detector

B. K-means clustering

Once we obtain the histograms for each image using the above procedure, categorization is implemented through unsupervised k-means clustering. The k-means clustering procedure partitions the data into k groups based on a predefined distance metric. The k-means procedure for partitioning the data is as follows:

- K initial cluster centroids are selected randomly $\{c_1, c_2, \dots c_k\}$
- Each point x_i in the dataset is assigned a cluster based on the nearest neighbouring cluster centroid, using a distance metric $D(x_i, c_k)$
- Each of the cluster centroids are recomputed as the mean of the points assigned to that cluster.

The last two steps are repeated until the cluster centroids converge. In the current model, since the data points are in form of histograms, the distance metric used was as follows:

$$D(x_i, c_k) = 1 - correlation(x_i, c_k)$$

$$= 1 - \frac{(x_i - \overline{x_i})(c_k - \overline{c_k})'}{\sqrt{(x_i - \overline{x_i})(x_i - \overline{x_i})'}\sqrt{(c_k - \overline{c_k})(c_k - \overline{c_k})'}}$$

Where $\overline{x_i}, \overline{c_k}$ represent the mean values of the two vectors.

To implement the three different types of categorizations relevant to the current study, the parameter *k* was varied accordingly. The clustering procedure was repeated three times – once each for super-ordinate, basic and subordinate level categories and the performance of the three was compared using certain performance metrics. So if the dataset has 8 sub-ordinate categories, which can be combined to form 4 basic and 2 super-ordinate categories, the k-means procedure was repeated thrice by setting k=8,4, and 2 respectively.

C. *Performance metrics*

Four performance indices were used to evaluate the clusters formed using the k-means procedure above and compare across different types of categorizations. Let $\Omega = \{\omega_1, \omega_2, \dots \omega_k\}$ be the set of clusters and $\mathbb{C} = \{c_1, c_2, \dots c_k\}$ be the set of true classes, and let N be the total number of points. Note that we assume the number of classes equal to the number of clusters, i.e. when super-ordinate categorization is being studied we assume the true number of underlying classes as 2, and so on. Following are the performance indices evaluated:

- **Rand Index –** The rand index treats each assignment as a series of decisions such that in total there would be N*(N-1)/2 decisions. A *true-positive* occurs when two points belong to the same class and cluster, *true-negative* occurs when the two points belong to different classes and clusters, *false-positive* occurs when two points belong to different classes but the same cluster, and *false-negative* occurs when the two points belong to the same class but different clusters. The Rand Index (RI) is then computed as follows:

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Normalized Mutual Information –** This index computes the information-theoretic mutual information between the classes and the clusters and normalizes by dividing by the sum of the entropies of the classes and the clusters:

$$NMI(\Omega; \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{(H(\mathbb{C}) + H(\Omega))/2}$$

$$I(\Omega;\mathbb{C}) = \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|}$$

$H(\mathbb{C})$ and $H(\Omega)$ are the entropies of the classes and clusters.

- **Purity –** This performance index assigns each class a label based on the majority elements in the class and finds the ratio of the correctly assigned data points:

$$purity(\Omega,\mathbb{C}) = \frac{1}{N}\sum_k \max_j |\omega_k \cap c_j|$$

- **Silhouette Index –** The silhouette index does not take into account the true classes of the points, it simply computes how well separated and compact the clusters are:

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}}$$

Where $a(i)$ – average dissimilarity of i-th object from all other objects in the same clusters and $b(i)$ –average dissimilarity of i-th object from all other objects in other clusters. The overall silhouette index is simply the average of the individual silhouette indices.

Hence, while the first three indices compute how well the clusters formed correspond to the actual classes within the data, the fourth index (silhouette) computes how compact and well separated the clusters are.

4. Dataset and Results

The dataset was constructed using Google Image search. 8 subordinate categories were taken (*Dalmatian, Foxhound, Crow, Pigeon, Bar stool, Rocking chair, Picnic table, and Coffee table*), which combined into 4 basic level (*dog, bird, chair and table*) and 2 super-ordinate level (*animal and furniture*) categories. 30 images for each subordinate category were taken. Figure 2 shows the hierarchy of these categorizations.

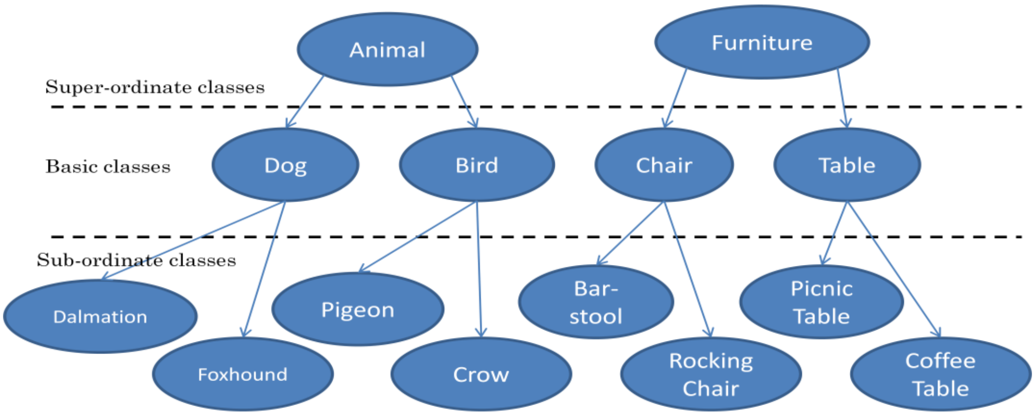
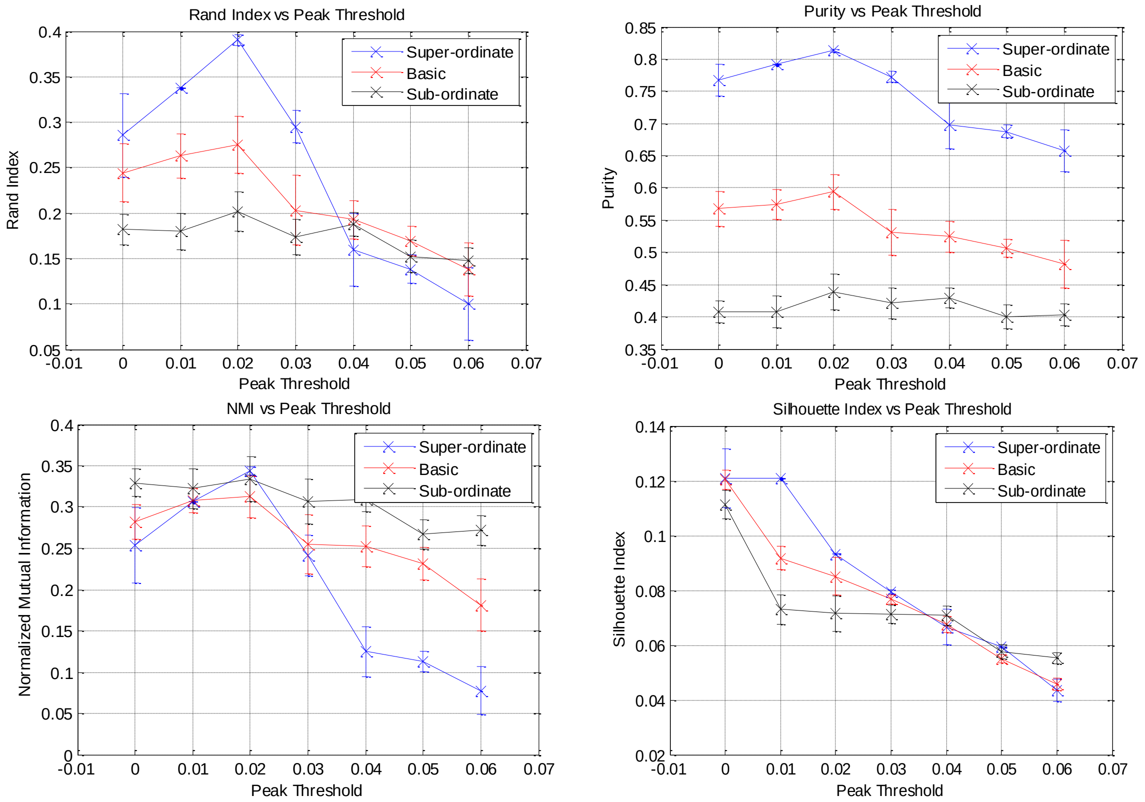


Figure 2, Hierarchy of the taxonomy used. 30 images for each subordinate category were used, giving a total of 240 images in the taxonomy

Three tests were performed on the data – performance variation with the number of key-points detected per image, performance variation with number of images per category, and role of expertise in basic v subordinate categorization.

A. Performance variation with number of key-points

The *Peak Threshold* parameter in implementing bag-of-features controls the number of key-points detected per image. Basically, the key-points with peak Difference of Gaussian (DoG) value less than the peak threshold are filtered out, [10]. Hence, the higher the value of this peak threshold, the lower the number of key-points detected. Figures 3a, 3b, 3c and 3d show the performance indices of the three types of categorizations as the *Peak Threshold* (PT) is varied.



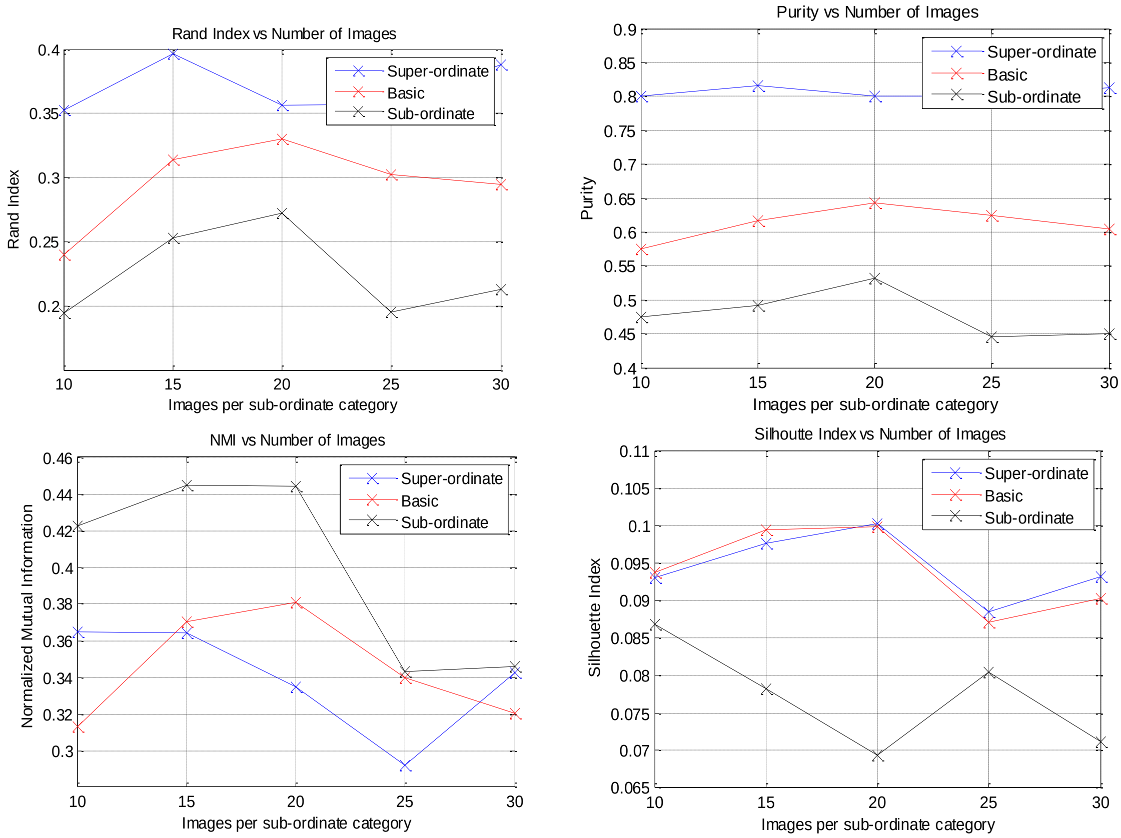
Figures 3a-3d, Rand index (3a), Purity (3b), Normalized Mutual Information (3c) and Silhouette Index (3d) against peak threshold for the three types of categorizations

Two important observations arise out of the above analysis. First, for low values of PT *super-ordinate* categorization dominates. This supports the hypothesis of the Marc and Joebert, that when only visual features are considered super-ordinate categories dominate. However for the normalized mutual information index the trend is reversed – subordinate categories dominate while the super-ordinate categories are lowest. This can be attributed to the preference of mutual information for higher number of cluster.

Second observation is that as the value of PT increases the performance of super-ordinate categorization decreases rapidly. While this is a surprising result given that as the number of key-points decreases the algorithm is expected to make distinctions between coarse classes (animal v furniture) better than fine classes (Dalmatian v Foxhound), the performance of the algorithm is quite heavily reduced and there isn't much significance in the results.

B. Performance variation with number of images

The hypothesis for this test was that as the number of images is increased the dominance should shift from the super-ordinate to perhaps basic of subordinate categories. Figures 4a-4d show the best performance indices (over 100 repetitions) of the three types of categorizations as the number of images per subordinate category is increased.

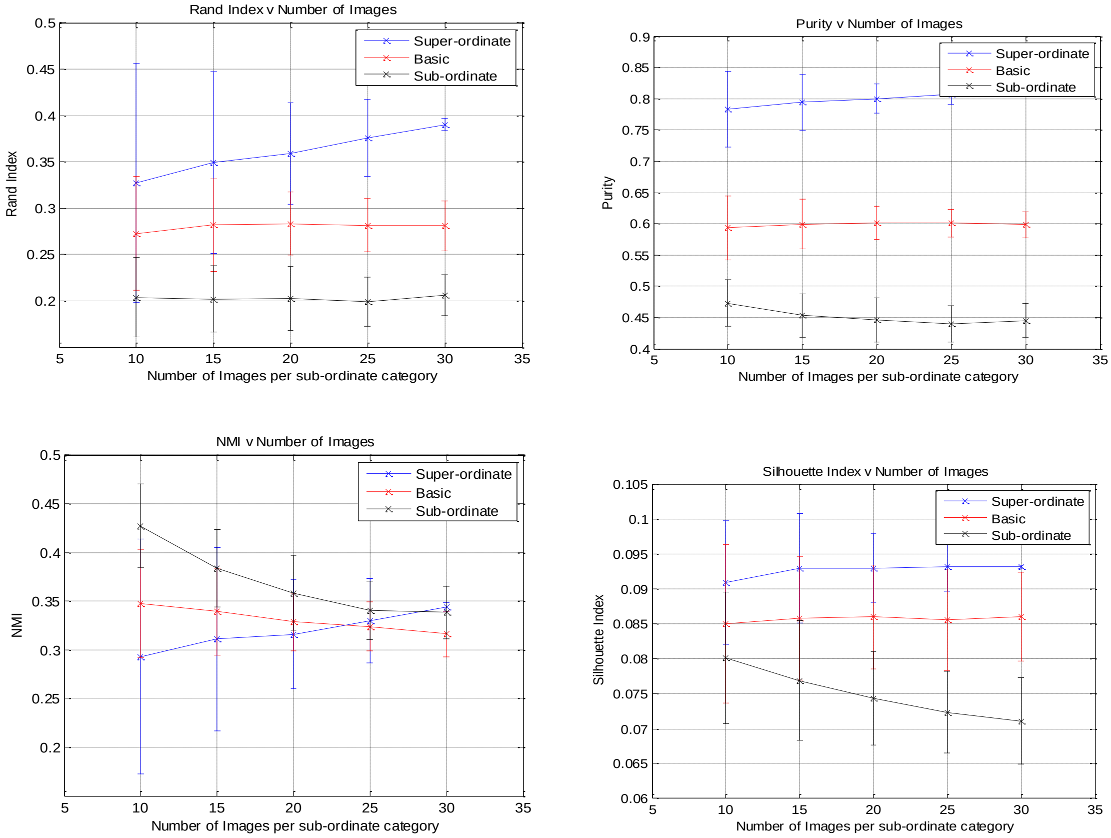


Figures 4a-4d, Rand index (3a), Purity (3b), Normalized Mutual Information (3c) and Silhouette Index (3d) against number of images per subordinate category for the three types of categorizations (best)

The results do not show much significance since the same category dominates as the number of images is varied. This can be attributed to perhaps the small increase in the number of images (10-30) whereas more insightful results might require an exponential increase in the number of images. A further refinement was made to the test to compute the *average* performance index over 100 repetitions, for the three types of categorizations as the number of images per sub-ordinate category was increased. Figures 5a-5d shows the results.

A. Role of expertise

This test was performed taking two basic categories at a time, say *dog* and *bird.* It is widely accepted in literature, [5], that for an expert in a certain domain (i.e. the person is exposed to many examples of that domain) what appear as subordinate level categories to most people appear as basic level categories to the expert. To simulate this, two clusterings were performed – once with an equal number of images for both basic categories and then with the number of images for one basic category double that of the other. Both the times three clusters were formed. So if we want to test expertise for *dog* category, we initially take 30 images each for *dog* and *bird,* cluster into 3 groups ({Dalmatian, Foxhound, Bird}) followed by taking 60 images for dog and 30 for bird and clustering into same three groups again. The performance increase or decrease was studied in terms of the rand index and is shown in Figure 5.



Figures 5a-5d, Rand index (3a), Purity (3b), Normalized Mutual Information (3c) and Silhouette Index (3d) against number of images per subordinate category for the three types of categorizations (average)

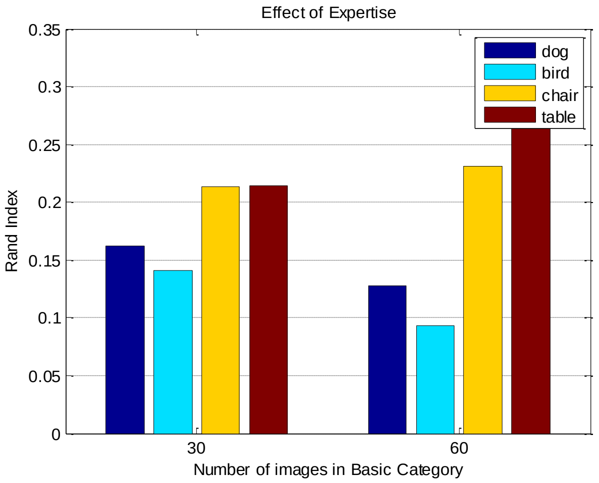


Figure 5, Effect of increasing the number of images of a basic category

It is evident from the above figure that the hypothesis is not supported. While the performance for the *chair* and *table* categories does increase, the same is not true for the *dog* and *bird* categories. One possible explanation for this is that in the visual feature space, these categories have very high variances and overlap significantly with each other. In such a case adding the number of training images need not show a performance increase necessarily. Another possible explanation is that the increase in number of images in not significant enough, and an exponential increase might be needed.

5. Discussion and Conclusions

Two hypotheses were tested in the current study – the dominance of super-ordinate categories when only visual features are processed, and the shift to subordinate categories as expertise in a domain increases. For the first hypothesis, the three different types of categorizations were studied using k-means clustering on a bag-of-features model. The performance metrics evaluated indeed show a preference for the super-ordinate categories, except for Normalized Mutual Information (NMI). This is possibly due to the volatile nature of NMI (slight changes in the clusters lead to very different values of NMI) coupled with its preference for a larger number of clusters. On the contrary, how much these indices are affected by the increasing difficulty of subordinate categorization as compared to super-ordinate or basic categorization is also not clear at the moment.

While, the results are somewhat supportive of the first hypothesis, that is not the case for the second hypothesis. Increase in the number of training images does not, in general, lead to a better performance for the subordinate case. While this may be due to the increase in the number of images not being sufficient enough in experiment (an exponential increase might be more relevant), it could also be attributed to the high variance in the classes in the visual feature domain. When there is a large overlap between the classes being separated, adding the number of images does not necessarily lead to a performance increase.

There were some general problems faced by the model. The most significant of these was that images belonging to the same class, but having significantly different backgrounds were most often put in separate clusters. This obviously leads to less accurate results. A possible solution to this would be to perform some kind of foreground extraction before applying the model. Another issue is the effect of the distance metric used. Correlation does provide an intuitive measure of similarity between histograms; however different metrics lead to different performances. It has been documented in literature that the Earth Movers Distance (EMD) or the Wasserstein metric performs best for comparison between histograms [11], but it was not used in the current model due to the extremely high computation time it requires. Also the overall performance of unsupervised clustering of objects is significantly lower than the corresponding human accuracy, even with a sophisticated model like this. Hence, the results obtained do not necessarily map to human thinking.

In conclusion, there is evidence for the dominance of the super-ordinate categorization in a visual processing task, however further analysis needs to be done before concrete claims can be made. The anomalous case of NMI needs to be studied in detail. Also the number of images per subordinate category needs to be increased further (about 150 each) and a more concrete test for the role of expertise needs to be devised. The validity of each of the performance metrics also needs to be addressed.

Acknowledgement

I would like to thank Dr. Amitabha Mukherjee for his guidance and support throughout the project.

References

[1] Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4, 328-350*.

[2] Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8, 382-439.*

[3] Murphy, G.L., Wisniewski, E.J. (1989). Categorizing objects in isolation and in scenes: what a superordinate is good for. *Journal of Exp Psychol Learn Mem Cogn 15: 572–586.*

[4] Mandler, J.M., McDonough, L. (2000). Advancing downward to the basic level. *Jounal of Cognit Dev : 1: 379 –403.*

[5] Johnson, K.E., Mervis, C.B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Expert Psychology*. *126, 248-277.*

[6] Mervis, C.B., Cisafin, M.A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development 53.*

[7] Marc, J.M.M., Joubert, O.R., Nespoulous, J.L. & Fabre-Thorpe, M (2009). The time-course of visual categorizations: you spot the animal faster than the bird. *PLoS one 4(6), e5927. doi:10.1371/journal.pone.0005927.*

[8] Jiang, Y.G., Ngo, C.W., Yang, J. (2007). Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. *CIVR'07*, *494-501.*

[9] Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *Int. Journal on Computer Vision, 2, 91-110.*

[10] Vedaldi, A., Fulkerson, B. (2008). VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/, accessed Sep, 2011.

[11] Rubner, Y., Tomasi, C., Guibas, L.J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision 40(2), 99–121.*