T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	An effective use of adaptive combination of visual features to retrieve image semantics from a hierarchical image database
Authors	Shreelekha Pandey, Pritee Khanna, Haruo Yokota
Citation	Journal of Visual Communication and Image Representation, Vol. 30, pp. 136-152
Pub. date	2015, 7
DOI	http://dx.doi.org/10.1016/j.jvcir.2015.03.010
Creative Commons	See next page.
Note	This file is author (final) version.

Powered by T2R2 (Science Tokyo Research Repository)

License



Creative Commons: CC BY-NC-ND

*Manuscript
Click here to view linked References

An Effective Use of Adaptive Combination of Visual Features to Retrieve Image Semantics from a Hierarchical Image Database

Shreelekha Pandey[a], Pritee Khanna[a,*], Haruo Yokota[b]

[a] *PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Dumna Airport Road, Jabalpur 482-005 M.P. India*
[b] *Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-kuTokyo, 152-8552 Japan*

Abstract

Correlating semantic and visual similarity of an image is a challenging task. Unlimited possibilities of objects classification in real world are challenges for learning based techniques. Semantics based categorization of images gives a semantically categorized hierarchical image database. This work utilizes the strength of such database and proposes a system for automatic semantics assignment to images using an adaptive combination of multiple visual features. 'Branch Selection Algorithm' selects only a few subtrees to search from this image database. Pruning Algorithms further reduce this search space. Correlation of semantic and visual similarities is also explored to understand overlapping of semantics in visual space. The efficacy of the proposed algorithms analyzed on hierarchical and non-hierarchical databases shows that the system is capable of assigning accurate general and specific semantics to images automatically.

Keywords: Image semantics, Semantic and visual similarities, Hierarchical image database, Search space pruning

1. Introduction

A person can easily infer some semantics from an image. For example, it is easy for us to infer semantics like *Sledding*, *Sports*, *Person*, *Grandparent and grandchildren* and many more from the image given in Fig. 1(a). We are adept to correlate visual similarity to semantic similarity and have natural instinct to group 'similar objects' in categories, and 'similar categories' to 'super-categories' [1]. As a result, for the image in Fig. 1(b), one can easily infer the semantic *Tiger*, followed by *Carnivore* and *Animal*. However, it is difficult for a computer

*Corresponding author
Email addresses: shreelekha@iiitdmj.ac.in (Shreelekha Pandey), pkhanna@iiitdmj.ac.in (Pritee Khanna), yokota@cs.titech.ac.jp (Haruo Yokota)

to infer such semantics from an image file. A computer easily computes low level features based on color, texture, and shape. Fig. 2 shows some images with their low level color features and high level semantics. Content Based Image Retrieval (CBIR) systems try to emulate human vision through visual similarity obtained in terms of low level image features to interpret images [2, 3]. The lack of coincidence between low-level visual data and high level semantics of images is known as semantic gap [4]. Development of universally acceptable algorithms to reduce semantic gap and characterize human vision for object recognition and image retrieval are in progress [5].

2. Problem Statement

Empowering computers to distinguish object categories in visual as well as in semantic space, is a challenging task. Obtaining knowledge of specific semantics is not straightforward even for humans many times. Consider the sunflower images of seven categories, namely *Swamp*, *Common*, *Giant*, *Showy*, *Maximilian*, *Prairie*, and *Jerusalem* in Fig. 3. A semantically categorized hierarchical image database may help us to automatically derive these semantics. The semantic based categorization of images leads to a hierarchical tree structure having images of different categories at various levels. This categorization may help us to understand the correlation between visual features and semantic of categories (e.g. *Animal*, *Vegetable*, *Fruit* etc.), which may further be utilized to provide specific semantics of the image. In an attempt to correlate visual and semantic similarities, this work aims to derive as exact semantics as possible at a moderate search cost by exploring only some branches of image tree.

3. Related Work

Learning algorithms for limited number of concepts are extensively used on flat image databases to reduce semantic gap [6, 7]. In a statistical modeling approach for automatic linguistic indexing of pictures, each of the 600 concepts is represented by a two-dimensional multi-resolution hidden Markov model and is trained with categorized images [6]. Generative probabilistic models for 101 object categories are learned through Bayesian incremental algorithm using a few training images for quick learning [7]. Visual recognition from semantic segmentation of photographs is learned to achieve 70.5% region-based recognition accuracy on a 21-class database [8]. Gaussian Mixture Models learned from bags of localized features of images with common semantic label are pooled into a density estimate for the corresponding semantic class [9]. Using the class densities, a minimum probability of error rule is used for image annotation and retrieval. For the same purpose, ExpectationMaximization algorithm and asymmetric PLSA (Probabilistic Latent Semantic Analysis) learning based on textual/visual information of images are also used to learn a model [10].

Optimization and estimation techniques are used in an automatic real-time image annotation system to represent objects by bags of weighted vectors grouped

on D2-clustering [11]. Hypothetical Local Mapping is utilized to develop a generalized mixture modeling technique for non-vector data. In another real-time image annotation approach key phrases of similar images are mined for candidate annotations [12]. The approach is scalable and robust to outliers as it does not require any training data. Image metadata and parametric dimensions were used to obtain a set of rules in a decision tree based automatic semantic annotation approach [13]. The system is developed with 3,231 manually labeled images and tested on 1,00,000+ Web images outside the training database.

Label correlations are explored to develop a two-dimensional active learner for image classification, and an adaptation algorithm is used to update the model [14]. Another label transfer based nonparametric system used a SIFT flow algorithm to retrieve dense scene correspondences from a fully annotated large database [15]. These correspondences are used to integrate multiple cues to recognize query images. A generic multiview latent space Markov network developed to relate image features and abstract concepts maximizes the likelihood of multiview data and minimizes a prediction loss on the labels from side information [16]. An optimal Image-tag relation matrix consistent to the observed tags and the visual similarity is obtained through a semi-supervised algorithm [17]. In another approach, an automatic news image caption generation system learned extractive and abstractive surface realization models from weakly labeled data in an unsupervised fashion [18].

Unlike traditional hypergraph learning, weights of hyperedges are adaptively learned in many works to improve the performance [19, 20, 21, 22]. The size of neighborhood is varied to generate a set of hyperedges, where weights are optimized by means of a regularizer [19]. Click data is integrated with the system to reduce the semantic gap [19, 21]. The images are also classified by combining information from labelled views [22]. A probability distribution constructed using high-order relationship is estimated through hypergraph. Also, visual and textual information are utilized for social image search [23]. The weights of hyperedges are learned to enhance the effects of informative visual words and tags. Learning employs a set of pseudo-relevant samples based on tags. Recently, The generative approach to identify visual neighborhood in training image set, is refined adaptively by discriminative hyperplane tree classifier [24].

A few more techniques exist in literature but unrestricted concepts in the real world limit the power of learning based approaches in general. Hierarchical structures are also used for the purpose of image retrieval [25], object recognition [26, 27], indexing [28] or codebook generation [29]. A tree structure, built by identifying various objects in the set of training images and arranging them on different levels depending on the relationship among the identied objects, is explored to get the desired results. Instead of focusing on learning techniques, Khanna et al. used a hierarchical image database to assign efficient semantics to a given image [30]. With the aim of deriving image semantics, a large hierarchical image database is used to establish a correlation between visual and semantic similarities. The present work extends the concept and aims to derive semantics with high precision in the reduced search cost.

The rest of the paper is organized as follows. Section 4 gives an insight

of related image databases. Proposed methodology is explained in Section 5. Section 6 discusses the representation of semantic categories of images in visual space. 'Branch Selection Algorithm' given in Section 7, is followed by pruning approaches in Section 8. Section 9 summarizes results and discussion on related issues. Finally, Section 10 concludes the work along with future directions.

4. Image Databases

The nature of image database influences the design and performance of semantics retrieval algorithms. For decades, researchers used self-collected images to show their results. Later, many domain specific databases having thousands of uncategorized images, e.g., WANG, UW, IRMA 10000, ZuBuD, and UCID came into existence [31, 2]. With large number of images, Caltech 101/256 [7], and MSRC [8] are some more challenging databases. Corel is a propriety database containing large number of pre-classified images ranging from *Animals* to *Outdoor Sports* but images with similar content are divided into different categories [3]. TinyImage consists of 80 million low resolution and noisy images based on WordNet synsets. ESP consists of millions of images labeled through an online game which results in an unbalanced image distribution across the semantic hierarchy. LabelMe (30K images) and Lotus Hill (50K images) provide labeled and segmented images of around 200 categories. A publicly available image database, ImageNet inherits hierarchical semantic structure from WordNet [32]. For each of its synset, images corresponding to WordNet synonyms are collected from several image search engines. ImageNet 2011 Winter Release has more than 14,000K images for nearly all object classes categorized by 21,841 synsets. ImageNet is developed to test the correspondence between cognitive and visual hierarchy which makes it a suitable candidate for experiments in this work. Fig. 4 shows some images from ImageNet.

5. Methodology

The proposed work is an attempt to establish a correspondence between semantic and visual hierarchy. The process flow of the system is shown in Fig. 5(a). Each node of a hierarchical image database represents a semantic category. Linear traversal of this large hierarchy to find the semantics of a query image is time consuming as well as error prone. Identification of the relevant subtrees for search would lead to efficient semantic retrieval. The process of assigning semantics to query image in this work is based on the visual signatures attached to nodes. Corresponding to each semantic category (node) in the hierarchy, a visual signature is generated through an offline procedure. A general as well as specific semantic is assigned to each query image through an online and automatic procedure.

Let the hierarchical image database has N subtrees at the first level. Based on the distance between the query image and visual signatures of these N subtrees, a Branch Selection algorithm selects only n branches ($< N$) for search.

The process is repeated at each level of hierarchy. The search cost is further reduced by pruning the search space returned by Branch Selection algorithm. Irrelevant subtrees are pruned by pruning approaches spread over two levels as shown in Fig. 5(b). The resulting compact search space is linearly traversed to assign semantics to the query image. Semantics of the nodes closer to the query image are assigned to it. The accuracy of semantics assigned by the proposed system also tells the extent to which low-level visual features and semantics of an image category maps to each other.

6. Representing Semantic Categories in Visual Space

Visual features of semantically similar images of a node are exploited to assign a visual signature to this node which represents this semantic in visual space. 'Branch Selection' and 'Pruning Approaches' make extensive use of the distinctive visual signatures of nodes to retain relevant subtrees.

6.1. Visual Feature Extraction

Some well-known color, texture, and shape features are used to compute visual signature of a node and to measure visual similarity of images [33].

6.1.1. Color Features

Color shows the strongest similarity to human eye [34]. Color histogram and moments in HSV color space are utilized in this work [2, 35, 36, 37]. Color features are extracted at global as well as local level considering five image regions (central ellipsoidal region and four surrounding regions) [37]. Quantized values of Hue and intensity give 68 dimensional Global Color Histogram (GCH). Histograms corresponding to five image regions are concatenated to form a Local Color Histogram (LCH). In general, GCH and LCH are represented as given in Eq. (1), where h_k^j and i_k^j are hue and intensity at k^{th} bin of j^{th} region. Eq. (1) represents GCH and LCH, where h_k^j and i_k^j are hue and intensity at k^{th} bin of j^{th} region. Similar to histogram, average (E), variance (σ) and skewness (s) of each channel of an image are used to obtain a 9 dimensional Global Color Moment (GCM) and a 45 dimensional Local Color Moment (LCM) as given in Eq. (2). Here j and k in $E_k^j/\sigma_k^j/s_k^j$ represent k^{th} color channel of j^{th} region.

$$
\begin{aligned}
F_{GCH} &= (h_1^1, h_2^1, ..., h_{51}^1, i_{52}^1, ..., i_{68}^1). \\
F_{LCH} &= (h_1^1, ..., h_{51}^1, i_{52}^1, ..., i_{68}^1, ..., h_1^5, ..., i_{68}^5). \qquad (1) \\
F_{GCM} &= \left(E_1^1, \sigma_1^1, s_1^1, E_2^1, \sigma_2^1, s_2^1, E_3^1, \sigma_3^1, s_3^1\right) \\
F_{LCM} &= \left(E_1^1, ..., s_3^1, ..., E_1^5, ..., s_3^5\right) \qquad (2)
\end{aligned}
$$

6.1.2. Texture Features

An image may contain textures of different degrees of detail and can be analyzed either on single or multiple scales. Grey level co-occurrence matrices, Tamura Features, structure elements, Discrete Cosine Transform, Wavelet Transforms, Gabor filters, and ICA Filters are popular texture features [34, 38]. The most frequently used Gabor filter as given by Eq. (3) is constructed using the Gabor function $g(x, y)$ as the mother wavelet. Suitable dilations and rotations of $g(x, y)$ through the generating function g_{mn} give a self-similar filter dictionary. Here $\theta = n\pi/K$, K = total number of orientations, S = number of scale, $a = (U_h/U_l)$-$1/(S$-$1)$. U_h and U_l are upper and lower centre frequencies of interest [39]. Gabor filter with four scales and six orientations are used here. Mean μ_{mn} and standard deviation σ_{mn} of the magnitude of wavelet transform coefficients give 48 dimensional Gabor Texture (GT) feature as given in Eq. (4).

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi iWx\right] \tag{3}$$

$$\begin{aligned} g_{mn}(x,y) &= a^{-m} g(x',y');\ m,n = int,\ m = 0,1,...,S-1 \\ x' &= a^{-m}(x\cos\theta + y\sin\theta), y' = a^{-m}(-x\sin\theta + y\cos\theta) \\ F_{GT} &= (\mu_i, \sigma_i),\ i = 1,2,3,...,24. \end{aligned} \tag{4}$$

6.1.3. Shape Features

Generic Fourier Descriptors, Zernike and Pseudo Zernike Moments, and Wavelet Descriptors are some popular shape representations [34]. Scale Invariant Feature Transform (SIFT) is frequently used to represent local image parts. SIFT extracts large number of keypoints from image that leads to robustness in extracting small objects among clutter [40]. This work uses SIFT with 4 octaves and 5 levels. K-means clustering forms 32 clusters per image [2]. For each cluster, count, mean, and variance gives SIFT Shape (SS) feature as given in Eq. (5).

$$F_{SS} = (CV_1,...,CV_{32}), (MV_1,...,MV_{32}), (VV_1,...,VV_{32}). \tag{5}$$

6.2. Computation of Visual Signature of a Node

GCH, LCH, GCM, LCM, GT and SS features are combined to represent an image. Each node in the image tree is assigned a visual signature by exploring color, texture, and shape features of all the images it contains. These visual signatures are low-level representatives of image semantics corresponding to respective nodes. Mean provides the closest prediction of any value in the data set, and therefore mean feature vector of all the images in a node is used to form its visual signatures.

Table 1: Summary of image features, visual signatures and similarity measures.

Features	Visual Feature (Image)	Visual Signature (Node)	Similarity Measure
Color (HSV space)	GCH: Global Color Histogram LCH: Local Color Histogram	GCHmean LCHmean	Vector Cosine [37]
	GCM: Global Central Moments LCM: Local Central Moments	GCMmean LCMmean	City Block [36]
Texture	GT: Gabor Texture	GTmean	Euclidean [39]
Shape	SS: SIFT Keypoints	SSmean	Earth Mover's [41]

6.3. Similarity Measures

Similarity measure suitable to a particular feature, as mentioned in literature, has been used to get the distances between images and nodes. Visual features of images, visual signatures of nodes along with similarity measures used in the work are summarized in Table 1.

7. Branch Selection Algorithm

With the aim of reducing the search space, 'Branch Selection Algorithm' selects some promising nodes to be searched. The sum of distances based on GCH, LCH, GCM, and LCM is color distance. Distance based on GT is texture distance, and the sum of distances based on SIFT Mean and Variance is shape distance. At each level, three lists corresponding to color, texture and shape distances are prepared to select a few subtrees (n) out of N subtrees available as shown in Fig. 6. This process is recursively performed on n subtrees selected at each level till it reaches the leaf nodes of these subtrees.

For the example Query image "n00450866_898" from *Pony - Trekking* synset, the output of the algorithm for $n = 3$ is shown in Fig. 7. At the first level $N = 11$ subtrees are used for experimentation. The algorithm selects *Geological Formation*, *Tree*, and *Sports* semantics at the top level. At the subsequent levels, synsets of these three high level semantics are chosen to get the search tree. With $N = 11$ and $n = 3$, the search space reduces by 73% in terms of subtrees to be searched. In this case, the algorithm selects only 51 synsets out of the total 365 synsets in the image tree to find the semantics of the query image. Thus search space is reduced by 86.03% w.r.t. the number of synsets to be searched, still keeping the desired subtree in consideration. Performance of 'Branch Selection Algorithm' depends on the parameters n and N. The relationship between n (number of subtrees to be chosen at any level) and N (subtrees at the top level) is explored in Section 9.2.

8. Pruning Approaches

'Branch Selection Algorithm' selects some general semantics to be searched to get specific semantics of the query image. The number of possible specific semantics depends on the height, width, and synsets of the n subtrees chosen at each level. Discarding all irrelevant nodes from the search path would lead us to precise semantic in the lowest search cost. A *relevant node* is the one that lies on the path leading to the node containing images semantically similar to the query image, while an *irrelevant node* leads to either a different path in the same subtree or a different subtree. Although, it is not possible to discard all irrelevant nodes, but the concept of *strict* or *soft* pruning discussed in this section may help in retaining relevant nodes and discarding the irrelevant nodes of a selected subtree. *Strict pruning* prunes the entire subtree if its root is not found to be relevant, while *soft pruning* removes only the irrelevant node from the path and not the entire subtree following it. The children of this irrelevant node become the children of its parent. For the query image "n00450866_898", *strict pruning* shown in Fig. 8(a) loses the target synset *Pony Trekking* because its parent node *Riding* is not found relevant. A less restrictive *soft pruning* approach in Fig. 8(b) preserves the target synset even if its parent is pruned.

Although *strict pruning* approach results in a smaller search space but has more tendency to produce false negatives as compared to *soft pruning*. Reduction in search space at the cost of efficiency of the system is not acceptable and hence *soft pruning* approach is further explored to design various pruning approaches. Moreover, these subtrees are successively pruned in two steps; referred as 'First Level Pruning' and 'Second Level Pruning'. These approaches work on the distances corresponding to the 'adaptive' selection of visual features, i.e., a 'dominant visual feature'. The three cases discussed in 'Branch Selection Algorithm' driven by the frequency of appearance of subtrees in three lists help to select dominant visual feature.

8.1. First Level Pruning

'Dominant feature' of a subtree is the feature which gives top rank to this subtree. Dqs_i is the distance (already calculated) between query image and i^{th} node of the subtree (having N_s nodes) w.r.t. dominant feature. If a subtree is given top ranking by shape feature, then Dqs are SS based distances. $Dmean$ and $Dmed$ are mean and median of Dqs. Relevance of nodes is determined on the basis of the relationship of Dqs with $Dmean$ and $Dmed$.

In addition, Average Absolute Deviation (AAD) and Median Absolute Deviation (MAD) are also used to find relevant node. AAD does not square the distance from the mean, so it is less affected by extreme observations than are the variance and standard deviation. MAD is a variation of AAD that is even less affected by extremes in the tail because the data in the tails have less influence on the calculation of the median than they do on the mean [42]. Extended mean distance ($Dmeanx$) and extended median distance ($Dmedx$) are calculated as given in Eq. (6). In this case, relevancy is determined on the

Table 2: Selected Pruning Algorithms.

First Level	Second Level	Pruning Algorithms
	AND	OR-AND
OR	OR	OR-OR
	Extended AND	OR-Ext AND
Extended AND	Extended AND	Ext AND-Ext AND
OR		OR-Inc
Extended AND	Parent Child Relationship	Ext AND-Inc
Extended OR		Ext OR-Inc

basis of the relationship of Dqs with $Dmeanx$ and $Dmedx$. The four possible combinations shown as 'First Level Pruning' in Fig. 9 are exhaustively tested.

$$Dmeanx = Dmean + \sum_{i=1}^{N_s} \left(\left| Dqs_i - Dmean \right| / N_s \right)$$
$$Dmedx = Dmed + median \left(\left| Dqs_i - Dmed \right| \right) \quad (6)$$

8.2. Second Level Pruning

To make pruning more powerful, the following two strategies are applied separately on the subtrees obtained as output of the 'First Level Pruning'.

Strategy 1. : Repeat the pruning strategy applied at the first level.

Strategy 2. : Explore Parent-Child Relationship. At any level of the hierarchy if the distance of the query image increases with the child nodes as compared to the parent node then it indicates that child nodes are faraway from the query image and there is no point in keeping them in consideration. Based on this observation if the distance with child nodes increases for more than half of the features then the child node is removed for further consideration.

8.3. Pruning Algorithms

Fig. 9 shows all identified combinations for the pruning algorithms. The terminology chosen for naming these pruning algorithms depicts the relationship of Dqs with $Dmean$ and $Dmed$ for different levels separated by a '-'. An algorithm named 'AND-OR' means that at the first level Dqs is less than equal to both $Dmean$ and $Dmed$, while at the second level Dqs is less than equal to either one of them. The term 'Ext' added in the name of algorithm represents the relationship of Dqs with $Dmeanx$ and $Dmedx$ in the similar way.

The nodes retained by pruning algorithm are used to assign semantics to the query image. The greed of a small search space may result in the pruning of the target subtrees, and hence poor precision. The performance of pruning

algorithms is judged on two parameters: (a) the number of nodes retained to be searched to assign semantics; and (b) the ability to preserve the relevant nodes. Restrictive nature of 'AND' reduces search space significantly but results in poor precision and hence not explored further at the 'First Level Pruning'. 'OR' operator being less restrictive improves the precision and therefore its combination with 'AND', 'OR', and 'Ext AND' are worth considering. 'Ext AND' is more flexible than 'AND'; and in terms of increasing flexibility its combinations are 'Ext AND-AND', 'Ext AND-OR', and 'Ext AND-Ext AND'; and the most flexible in the group 'Ext AND-Ext AND' may maintain a good precision-pruning ratio and thus be a good candidate to explore. Being least restrictive, 'Ext OR' does not prune much search space and therefore all of its combinations are discarded using Strategy 1 at the second level. Pertaining to Strategy 2, all combinations are less restrictive giving good results except 'AND' and hence excluded. On the basis of these considerations and initial results, only seven pruning algorithms given in Table 2 are exhaustively tested.

9. Results and Discussion

Performance of the proposed algorithms is analyzed to better understand the correspondence between visual and semantic similarity in a semantically categorized hierarchical image database.

9.1. Experimental Setup

Experiments are performed on a system with Core 2 Quad processor, 8GB RAM and 500GB HDD. In lieu of the limited computing resources, a subset of ImageNet, with 3,32,000 images from 365 synsets belonging to the most common 11 categories, as shown in the Table 3 is used for experimentation. The experiments are also performed on WANG for comparison purpose. The proposed system is exhaustively tested with two sets of query images. Set-1 of query images consists of four subsets automatically formed by randomly selecting 5% of images from each synset of test database. Average performance of the system on these four subsets is considered to overcome any sort of bias. Set-2 of 36,500 query images is collected through Google image search engine [43], which consists of 100 images for each synset under the categories shown in Table 3. The system automatically assigns semantics to query images and checks it against its predefined semantics. No manual intervention is required during performance evaluation as human subjectivity may affect the understanding of correlation between visual and semantic similarity.

9.2. Performance of Branch Selection Algorithm

'Branch Selection Algorithm' definitely prunes the search space but there is a trade-off between speeding up the search process and obtaining accurate semantics. The performance of the algorithm greatly depends on the number of subtrees selected at each level, i.e., the value of parameter n. Choosing n too small as compared to N will reduce the search space considerably and speed

Table 3: A subset of ImageNet used for Experimentation.

Subtree	Width	Depth	# of Synsets	# of Images
Animal	9	9	32	38,000
Appliance	4	4	29	32,000
Fabric	2	5	12	11,500
Flower	9	3	24	26,000
Fruit	6	5	42	30,500
Geological Formation	5	5	50	55,000
Person	12	4	34	16,500
Sport, Athletic	5	4	23	30,500
Structure	6	6	36	33,000
Tree	7	6	42	24,000
Vegetable	6	5	41	35,000
Total (average of 910 images/synset)			**365**	**3,32,000**

up the process; but at the same time it would tend to decrease the precision of the system. Here, precision specifies selection of the target subtree for a query image. Precision improves with increasing value of n but enlarges the space to be searched and hence results in increased search time. 'Branch Selection Algorithm' is a flexible algorithm which can easily be fine-tuned as required. To understand this flexibility, experiments are performed on the database given in Table 3 with different values of n; and variations in precision as well as time are observed. Further, obtaining meaningful semantics of images with smaller value of n indicates a strong correlation between semantic and visual similarity.

The precision of the algorithm is directly related with the value of n chosen but the actual search time depends on the computing facility being used. Each graph in Fig. 10(a)-(k) depicts relative increase in percentage for precision and time with increasing value of n for query images from set-1. Total search space is represented in terms of N subtrees at the top level. With $N = 11$; $n = 1$ leads to exploration of only 9% of the search space, while $n = 7$, requires almost 64% of the search space to find semantics. In Fig. 10(a), for *Animal* hierarchy when n increases from 1 to 2, precision increases from 18.4% to 42.2% (129% $\uparrow$) and search time goes up by 95%. Increasing n from 2 to 3 elevates precision to 66.8% (58% $\uparrow$) but doubles the cost of searching. At $n = 4$ precision reaches to 82.8% (24% $\uparrow$) with 49% increase in search time. Fig. 10(a)-(k) reveals that initial increments in the value of n improves precision significantly but with drastic increase in search time. *Animal*, *Geological Formation* and *Vegetable* subtrees are exceptions where percentage increase in precision is more as compared to time when n changes from 1 to 2. It can also be noticed that precision and time follow increasing pattern with increasing value of n but the percentage increase in precision is quite less as compared to percentage increase in the search time for higher values of n. This is true for all 11 semantics.

To get the idea of overall performance of 'Branch selection Algorithm', the

average performance over all 11 subtrees is shown in Fig 10(l). This graph confirms that time increases more rapidly as compared to precision with increasing value of n. For n varying from 1 to 2, and 2 to 3, time increases by 111% and 84% respectively; but gain in efficiency drops from 69% to 29%. Increase in precision is more till n reaches 3. Beyond this, 8%-6% improvement in precision is observed with more than 30% increase in search time till n reaches 6. Thus $n = 3$ (73% pruning) and 4 (64% pruning) seems good options here which give considerable improvement in accuracy. These choices of n are further validated through Fig. 11.

As shown in Fig. 11(a) precision increases significantly up to $n = 3$ for all hierarchies. Also, hierarchies like *Animal* (24%), *Fabric* (16%), *Geological Formation* (12%), *Sports* (33%), and *Vegetable* (12%) report good increase in precision at $n = 4$. However, at $n = 7$ (44% pruning), most of the hierarchies report 90% precision with an overall precision of 89%. For $n > 7$, less than 40% space would be pruned and ultimately 'Branch Selection Algorithm' starts losing its benefit. Keeping it in consideration experiments are not performed for higher values of n. Fig. 11(b) shows that for $n = 1$ and $n = 2$ (more than 90% pruning of the actual search space in terms of subtrees), algorithm achieves 50% (approx.) precision only as it often rejects relevant subtree and therefore it would be a futile exercise to generate further results on its output. Considerable improvement in accuracy is observed at $n = 3$ (87% pruning, 69% precision) and $n = 4$ (78% pruning, 74% precision). Fig. 11(b) also shows that with each increasing value of n search space increases substantially. The least increment of 4% in search space is reported when n increases from 1 to 2. The other receptive increments in search space are 8%, 9%, 12%, 12%, and 10% (i.e., approximately 10% every time). But growth in precision (22% and 16%) are notable when n increases from 1 to 2 and from 2 to 3, respectively. After this, precision increases constantly by 5%. This shows that the benefit over precision is lost with increasing value of n. Based on this experimental analysis, it is concluded that $n = \lceil N/3 \rceil$ or $\lceil N/4 \rceil$ is suitable to derive semantics with good precision in the reduced search cost. On the basis of these observations, suggested pruning range is 65% - 75%. Although the experiments with query images in set-1 reports the pruning of approximately 80% for $n = 3$ or 4, which is a little more than what is actually perceived. Fig. 11(b) also implies that one may opt for higher values of n to improve precision and a still small search space.

Following the suggested range, the performance of 'Branch Selection Algorithm' at $n = 3$ and $n = 4$ is compared for query images from set-1 (inside) and set-2 (outside) as shown in Fig. 12(a) and (b). At $n = 3$ average precision achieved for set-1 is 69%, while for set-2 is 65%. At $n = 4$, average precision is 74% for both sets of query images. For inside query images, *Fabric* and *Sports*; and for outside query images, *Fabric* are outliers as the precision achieved is only 40% approx. On excluding these outliers, average precision achieved for set-1 at $n = 3$ is 77%, while for set-2 is 67%. This difference in the performance for two sets of query images further decreases at $n = 4$. Excluding outliers at $n = 4$ gives 82% average precision for set-1 and 77% for set-2.

The discussion reveals that the system can be fine-tuned with 'Branch Selection Algorithm'. The execution time of the algorithm is significantly affected by the width and depth of the subtrees selected at level 1. If the width of subtrees is more, then it would lead to selection of more subtrees to be searched; while higher depth means more iteration. Moreover, actual pruning achieved in terms of number of synsets to be searched is much more than what is perceived in terms of subtrees selected. Examples given in Fig. 7 and Fig. 14(a) show the actual pruning achieved with 'Branch Selection Algorithm' at $n = 3$. Average execution time of the algorithm is 70 sec., which is high for an online search. However, considering the computational facility used and the absence of an appropriate indexing of image features with this size of database, results are encouraging. In the real time environment, the algorithm would be executed at the server end with indexed features substantially reducing its execution time.

9.3. Performance of Pruning Algorithms

In order to put a strict check on the performance, pruning algorithms are applied on the output of the 'Branch Selection Algorithm' at $n = \lceil N/4 \rceil$. Pruning algorithms must maintain the precision obtained through 'Branch Selection Algorithm' and also prune as many irrelevant nodes as possible. The inverse relationship of pruning percentage and retained precision is clearly visible in Fig 13. With 'First Level Pruning' given in Fig. 13(a), OR approach prunes search space by 41% at the cost of losing the precision further by 13%. The pruning gets lower with 'Ext AND' (26%) and 'Ext OR' (17%) and hence precision is better maintained (95% and 97% respectively). With 'Second Level Pruning' shown in Fig. 13(b), in the group of Strategy 1 'OR-AND' gives highest pruning (73%) but retains lowest precision (56%). 'OR-OR' prunes 67% space and retains 65% precision. 'OR-Ext AND' retains a better precision (77%) but with more cost of searching. Finally, 'Ext AND-Ext AND' in this group retains 87% precision with 44% pruning. As stated earlier, AND/OR strategy followed at any level of pruning has a strong impact on pruning percentage and precision. In the group of Strategy 2, all pruning algorithms provide accuracy in the range of 87% to 97% which is quite acceptable looking at their pruning percentage (67% to 58%). Instead of checking the goodness of a node against a fixed boundary framed by some predefined criteria in Strategy 1, exploring parent-child relationship in Strategy 2 enables them to perform better. Execution time of pruning algorithms depends on the height and width of the subtrees selected by 'Branch Selection Algorithm'. Average execution time of Strategy 1 and Strategy 2 algorithms are 0.157 sec. and 0.155 sec., respectively.

9.4. A Case Study

Output of 'Branch Selection Algorithm' @$n = 3$ for the query image *Fruit* $\rightarrow$ *Edible Fruit* $\rightarrow$ *Citrus Fruit* $\rightarrow$ *Orange* $\rightarrow$ "n07747607_21333" is shown in Fig. 14(a). The algorithm successfully retrieves the desired *Fruit* synset. It also reduces the actual search space (11 subtrees, 365 synsets) to only 3 subtrees with 54 synsets. Fig. 14(b) and (c) show the search space after first and second level

of pruning. 'First Level Pruning' retains 43 synsets which is further reduced to 22 by 'Second Level Pruning'. Finally, these 22 synsets are arranged in the increasing order of distances from the query image and the top three synsets are selected to generate the suggested list of semantics as shown in Table 4.

9.5. Semantics assigned to Query Images

For some of the query images, Table 4 summarizes the general and specific semantics derived through 'Branch Selection Algorithm' @$n = 3$ followed by 'Ext OR-Inc' pruning. Consider the images from any category; say *Animal* for example, then the retrieved semantics corresponds not only to *Animal* (in general) but also to a number of specific semantics like *Yearling*. The system successfully retrieves a specific semantic *Riding* which is relevant enough to the first query image irrespective of the fact that *Riding* belongs to a general semantic *Sports* in the test database. Assigning very specific semantic is difficult for a common user but the proposed system makes this task not only easy but also automatic. The system does not require user intervention or feedback at any stage to get a specific semantic for an image.

9.6. Correlation between Semantic and Visual Similarity

Table 5 gives a confusion matrix obtained through the proposed system at $n = 3$ for set-1 of query images. Diagonal values give precision obtained for each category and the remaining values show the overlapping of categories in visual space. Higher precision obtained at smaller n characterizes the existence of strong correlation between the semantics and visual similarity. The system out-performs for *Appliance* (94% precision), but under-performs for *Fabric* and *Sports* (approx. 30% precision). This happens due to the nature of images constituting these categories. The images are analyzed to understand this association; and misclassifications are explained through some characteristic images from various categories as given in Table 6-Table 8.

Most of the images in *Appliance* category have constant background with an object at the center. Black, white, and shades of gray are dominating this category. Shape features are also very prominent as objects have an attached geometry (mostly rectangular) with them. Images given in Table 6 show that *Appliance* synset contains images which are very close to each other visually as well as semantically, while *Fabric* or *Sports* consist of visually different images poorly related on the semantics. Although images in *Fabric* category are heterogeneous but the most of them have a constant background with a piece/pile of clothes at the foreground. Images of canvas of sail boat are kept in this category and as a result *Fabric* category overlaps the most 36% with *Appliance*. Color and Shape features of the two images from *Fabric* category in Table 7 are responsible for classifying them as *Appliance* by the proposed system. Similar is the reason behind classifying many similar images from other categories (like *Sports*) as *Appliance*. *Geological Formation* contains 6% images mapped to *Appliance*. These images are mostly of the places which are having a significant amount of white color, for ex., places covered with ice. Geometry is prominent

Table 4: Semantics assigned by the proposed system as compared to ImageNet semantic.
* Image database consists of 11 subtrees (general semantics) and 365 synsets (specific semantics).

ImageNet Semantics	Query Image	Proposed Semantics (General)	Synsets selected after Branch Selection	First Level Pruning	Second Level Pruning	Synsets Pruned (%)	Proposed Semantics (Specific)
Animal, Chordate, Vertebrate, Mammal, Placental, Yearling		Fabric, Sports, Animal	38	28	19	94.79	Yearling, Riding, Animal
Animal, Chordate, Vertebrate, Mammal, Placental, Carnivore, Bear, Brown Bear		Animal, Tree, Geo. Form.	59	41	26	92.88	Brown Bear, Natural Elevation, Carnivore
Appliance, Home appliance, Kitchen appliance, Coffee maker		Appliance	24	17	14	96.16	Espresso maker, Ice machine, Silex
Appliance, Home appliance, White goods, Refrigerator		Appliance	24	18	14	96.16	Ice machine, Refrigerator, Deep freeze
Fabric, Piece of cloth, Towel		Appliance, Person, Fabric	42	27	23	93.70	Hand towel, Towel, Paper towel
Fabric, Rayon, Acetate rayon		Fabric, Sports, Structure	47	33	21	94.25	Rayon, Towel, Viscos rayon
Flower, Marigold		Tree, Fruit, Flower	55	45	29	92.05	African marigold, French marigold, Golden wattle
Flower, Pink, Sweet William		Flower, Fruit, Vegetable	49	35	29	92.05	Fringed pink, Pink, Bing cherry
Fruit, Edible Fruit, Citrus Fruit, Orange		Fruit, Vegetable, Fabric	54	43	22	93.97	Orange, Jaffa orange, Citrus fruit
Fruit, Berry, Baneberry		Flower, Fruit, Vegetable	34	26	21	94.25	Baneberry, Cranberry, Currant
Geological formation, Slope, Hillside, Brae		Geo. Form., Vegetable, Tree	57	42	26	92.88	Down, Brae, Tableland

ImageNet Semantics	Query Image	Proposed Semantics (General)	Synsets selected after Branch Selection	First Level Pruning	Second Level Pruning	Synsets Pruned (%)	Proposed Semantics (Specific)
Geological formation, Shore, Strand		Geo. Form., Appliance, Fabric	51	39	24	93.42	Tideland, Landfall, Strand
Person, Planner, Schemer, Politician		Person, Appliance, Structure	42	30	14	96.16	Filer, Optimist, Litigant
Person		Person, Appliance, Flower	40	31	16	95.62	Appellant, Optimist, Filer
Sports, Skating		Person, Sports, Appliance	38	28	15	95.89	Figure skating, Skating, Person
Sports, Riding, Pony-trekking		Animal, Geo. Form., Sports	46	32	14	96.16	Pachyderm, Yearling, Pony-trekking
Structure, Tower, Bell Tower, Campanile		Appliance, Person, Structure	52	41	28	92.33	Bell tower, Campanile, Flophouse
Structure, Defensive structure, Stronghold, Bastian, Kremlin		Structure, Appliance, Person	55	42	27	92.60	Tower, Kremlin, Flophouse
Tree, Acacia, Wattle		Tree, Vegetable, Flower	46	31	23	93.70	Golden wattle, Acacia, Celery top pine
Tree, Gum tree, Eucalyptus		Geo. Form., Animal, Tree	59	44	20	94.52	Eucalyptus, Rose gum, Pinon, pinyon
Vegetable, Onion, Vidalia onion		Fruit, Appliance, Vegetable	64	47	36	90.14	Vidalia onion, Bermuda onion, Spanish onion
Vegetable, Cruciferous vegetable, Cabbage		Vegetable, Appliance, Fruit	59	48	28	92.33	Chinese cabbage, Cruciferous vegetable, Muskmelon

Table 5: Confusion matrix obtained with the proposed system for query images from set-1 ($N = 11$, $n = 3$).

	Animal	Appli-ance	Fabric	Flower	Fruit	Geo. Form.	Person	Sports	Struc-ture	Tree	Veget-able
Animal	**67**	5	2	1	2	2	7	1	2	8	3
Appliance	1	**94**	0	0	1	0	2	0	1	0	0
Fabric	4	36	**33**	1	2	1	11	1	8	1	1
Flower	2	2	1	**76**	2	1	4	1	1	8	2
Fruit	2	3	1	2	**81**	0	2	1	0	6	2
Geo. Form.	2	6	1	1	2	**74**	5	0	2	4	2
Person	2	4	2	2	3	0	**81**	2	1	1	3
Sports	9	16	2	2	3	5	18	**31**	7	3	3
Structure	2	6	1	0	2	1	5	1	**78**	3	1
Tree	3	2	1	1	2	2	3	1	1	**82**	2
Vegetable	6	9	3	2	4	2	10	2	1	6	**56**

Table 6: Some representative images of *Appliance*, *Fabric*, and *Sports* from ImageNet.



for the images belonging to *Structure* category which sometimes diverts these images (6%) to *Appliance*. *Vegetable* semantic is confused the most 10% with *Person* and 9% with *Appliance*. Correspondence with *Appliance* is due to the occurrence of an actual appliance or a single vegetable in the image.

Overlapping of *Animal*, *Flower*, *Fruit* and *Vegetable* with *Tree* is quite understandable. Correlation between *Animal* and *Tree* is quite obvious as most of the animal habitats are near greenery. Similarly, one can found fruits/flowers/ vegetables on plants and the database contains many images like this. Some images from these categories, as shown in Table 7, are classified as *Tree* seems to gel with the desired semantic. The images having significant amount of red/orange color are often put in *Flower*, *Fruit*, or *Vegetable* category. In the visual space, *Person* semantic overlaps with most of the semantics. In fact all other categories contain pictures with prominent human figure and hence as-

Table 7: A few images from ImageNet to understand the confusion matrix in Table 7.

* ImageNet semantics are in *italic*. Proposed semantics are in **bold**.

Animal		*Fabric*		*Flower*	
Fabric	**Tree**	**Appliance**	**Appliance**	**Tree**	**Tree**
					
Fruit		*Geological Formation*		*Person*	
Tree	**Tree**	**Appliance**	**Structure**	**Structure**	**Flower**, **Fruit**, **Vegetable**
					
Sports		*Structure*		*Tree*	
Appliance	**Appliance**, **Sports**	**Appliance**	**Appliance**	**Structure**	**Fruit**, **Flower**
					
		Vegetable			
		Appliance	**Appliance**		
					

Table 8: Example images from ImageNet identified as *Person* by the proposed system.

Animal	*Appliance*	*Fabric*
 	 	 
Flower	*Fruit*	*Geological Formation*
 	 	 
Sports	*Structure*	*Vegetable*
 	 	 

signed *Person* semantic by the proposed system as shown in Table 8. Many images in *Fabric* semantic contain persons wearing some dress material. Also appearance of person in *Sports* is obvious. As a result, system classifies 11% Fabric and 18% of *Sports* images as *Person*.

Although images in ImageNet are verified and labelled by humans using AMT (Amazon Mechanical Turk) platform and the difference in user judgment has also been taken care of. However, comparing the semantics assigned by the proposed system with that assigned by ImageNet as shown in Table 7 and Table 8 reveals that many times the proposed system is assigning more acceptable semantics. Present analysis is performed automatically and such images are considered as misclassification. Such misclassified images adversely affects the computation of visual signatures of semantics. The proposed approach may help to identify such cases and re-classification of these images would further improve the performance. Considering all such cases as true positive reveals a strong correlation between the semantic and visual similarity in a hierarchical image database.

9.7. Insertion Intensiveness of the System

Image features and visual signatures of nodes in the database are generated offline and new images are not inserted during experimentation. In real life, database may be kept in the updated mode. This insertion intensiveness can be easily handled online. Insertion of an image requires computation of its features and updating the visual signature of the node to which this image is added. The re-computation of visual signature is not computationally intensive.

9.8. Comparative Evaluations with WANG Database

ImageNet is chosen for experimentation due to its rich hierarchical structure. But most of the available image databases like WANG, Caltech 101/256, La-

Table 9: Average precision values obtained on WANG database.

Category	Proposed System ($n = 3$, $N = 10$)	[44]	[45]	[46]
Reduction in search space	70%	0%	0%	0%
Africa	0.93	1	0.76	1
Beach	0.90	0.58	0.587	1
Bus	0.96	0.61	0.963	1
Dinosaur	1.00	0.71	1	1
Elephant	0.96	0.49	0.741	1
Flower	0.97	0.58	0.945	1
Food	0.90	0.48	0.733	1
Horse	0.95	0.72	0.941	1
Monument	0.90	0.57	0.714	1
Natural Scene	0.95	0.47	0.457	1
Average	**0.942**	**0.621**	**0.7841**	**1**

belMe and others are flat in nature. In the absence of hierarchy, branch selection and pruning algorithms do not serve any purpose. WANG is the commonly used database, and hence arranged in ten categories at the top level for comparison. Table 9 compares performance of the proposed system with other related work on WANG. The proposed system gives an overall precision of 94.2% exploring only 30% of WANG as compared to 100% search done by other approaches. Table 10(a) shows that 63.6% average classification accuracy is obtained with an automatic linguistic indexing of pictures by a statistical modeling approach [6] on WANG. The proposed system reports 73.9% average precision at $n = 1$ (90% pruning), which reaches to 94.2% at $n = 3$ (still 70% pruning) on WANG as shown in Table 10(b).

Actual images in WANG are interpreted in the context of Table 10 to understand the overlapping of semantic categories in the visual space. The images in *Dinosaur* category form the most compact visual space, and hence there is no confusion for the proposed system. However, visual space of *Dinosaur* overlaps with 6 categories in [6]; as compared to only 1 category i.e., *Elephant* in the proposed system. Images in *Flower* have similar background with strong emphasis on color which makes their classification easy. However, in Table 10(a) visual features of 44 images overlap with *Flower*. This number is very less in Table 10(b) (19 images at $n = 1$ and 1 image at $n = 3$). The proposed system works much better for *Africa*, *Beach*, *Bus*, *Elephant*, and *Horse* which indicates their compact representations in the visual space. The proposed system confuses *Monument* with *Beach* at $n = 1$, but this confusion vanishes at $n = 3$. Images in *Beach* and *Natural Scene* are challenge for automatic classification. *Beach* images are subset of *Natural Scene* images and hence their overlapping in

Table 10: A comparison on WANG database using confusion matrix.

(a) Automatic image semantics obtained by [16]

Precision @ 1	Africa	Beach	Bus	Dinosaur	Elephant	Flower	Food	Horse	Monument	Natural Scene
Africa	**52**	2	0	8	16	10	2	0	4	6
Beach	0	**32**	0	0	0	2	0	2	6	58
Bus	0	18	**46**	2	8	0	4	0	6	16
Dinosaur	0	0	0	**100**	0	0	0	0	0	0
Elephant	8	0	0	8	**40**	0	0	8	2	34
Flower	0	0	0	0	0	**90**	6	0	2	2
Food	6	4	2	6	0	8	**68**	0	0	6
Horse	0	2	0	0	4	24	6	**60**	0	4
Monument	8	4	0	8	6	0	4	0	**64**	6
Natural Scene	0	6	0	2	2	0	0	0	6	**84**

(b) Automatic image semantics obtained through the proposed system at $N = 10$
(@1: $n = 1$, 90% reduction in search space; and @3: $n = 3$, 70% reduction in search space)

	Africa		Beach		Bus		Dinosaur		Elephant		Flower		Food		Horse		Monument		Natural	Scene
	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3
Africa	**60**	**95**	2	1	0	0	0	0	14	2	8	1	3	0	7	0	4	0	2	1
Beach	5	2	**69**	**91**	1	1	0	0	5	0	1	0	0	0	3	2	3	2	13	2
Bus	0	0	3	1	**85**	**96**	0	0	0	0	0	0	0	0	1	0	5	0	6	3
Dinosaur	0	0	0	0	0	0	**100**	**100**	0	0	0	0	0	0	0	0	0	0	0	0
Elephant	2	0	5	1	0	0	3	1	**70**	**96**	0	0	0	0	7	1	6	0	7	1
Flower	4	0	1	0	0	0	0	0	1	1	**92**	**97**	2	2	0	0	0	0	0	0
Food	5	1	2	0	1	1	0	0	12	5	9	0	**61**	**91**	6	0	1	0	3	2
Horse	6	1	0	0	0	0	0	0	7	1	0	0	0	0	**82**	**95**	5	3	0	0
Monument	12	0	21	0	0	0	0	0	6	1	1	0	3	1	4	5	**48**	**92**	5	1
Natural Scene	0	0	19	2	1	0	0	0	4	1	0	0	0	0	2	0	2	2	**72**	**95**

Table 11: Some images from WANG with different semantic assigned by the proposed system.

#	Image	WANG	Proposed	#	Image	WANG	Proposed
1		Africa	Natural Scene	6		Africa	Beach
2		Monuments	Natural Scene	7		Beach	Bus
3		Beach	Natural Scene	8		Beach	Monuments
4		Food	Natural Scene	9		Monuments	Elephant
5		Africa	Dinosaur	10		Monuments	Food

visual space is quite obvious that further proves the correspondence of semantic categories and visual features.

Table 11 helps to comprehend the situations in which the proposed system fails to get correct classification on WANG. The first four images are classified under *Natural Scene* by the proposed system and a closer examination shows that the assigned semantic suits well for the first three images. Green and Sand colors in the image 4 place it in *Natural Scene* and show lack of correspondence between semantic and visual similarity for this particular case. Due to closeness between the shape features of image 5 and *Dinosaur*, this image is classified as *Dinosaur*. For image 6 the assigned semantic *Beach* gels better with the visual content of the image. These images may belong to some scene from *Africa* but it is difficult for a common user to keep them in this category. Similarly for images 7, 8 and 9, a combination of color and texture features controls the semantic assignments, and hence *Bus*, *Monuments* and *Elephant* is assigned, respectively. Due to existence of a proper blend of red and yellow colors, image 10 lies in the visual space of *Food* semantics. The acquired knowledge from these discussions may be used to minimize this overlapping.

10. Conclusion and Future Scope

The experiments show that the adaptive combination of multiple low level image features serves well to assign semantic to any image. The performance of the system is quantized in terms of automatically retrieved semantics without any manual intervention. The results help to correlate visual and semantic similarities in a semantically categorized image database. The size of the visual

signature seems to be large for an online application, and proper indexing of visual signatures would help to reduce execution time.

Correlation among semantic categories is also explored and observations are acceptable for pairs like *Tree-Flower*, *Tree-Fruit*, etc. The larger correlation of *Fabric* and *Sports* with other categories depicts that representation of these categories is not compact in visual space, and hence overlaps with other semantics. Such associations can be reduced by incorporating the concepts of object extraction and clustering. Efforts are required to obtain more compact and efficient visual signatures of semantic categories.

In case 'Branch Selection Algorithm' fails to select proper subtree, user feedback may help to backtrack and select appropriate subtrees. Also, user can provide some keyword(s) to guide the search. For example, user may not be able to distinguish various categories of *Sunflower* but he can still divert the search to at least *Sunflower* or *Flower*.

Acknowledgements

This work is done during the JSPS Invitation Fellowship for Research in Japan (Long-Term) from May 2011 to Jan 2012. The authors acknowledge the support provided by JSPS Fellowship.

References

[1] R. J. Sternberg, Cognitive psychology , fifth ed., Wadsworth Cencage Learning, 2008.

[2] T. Deselaers, D. Keysers, H. Ney, Features for image retrieval: an experimental comparison, Information Retrieval 11 (2) (2008) 77–107.

[3] Y. Liu, D. Zhang, G. Lu, W.-Y. Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recognition 40 (1) (2007) 262–282.

[4] H. W. Hui, D. Mohamad, N. Ismail, Semantic gap in cbir: Automatic objects spatial relationships semantic extraction and representation, International Journal Of Image Processing (IJIP) 4 (3) (2010) 192–204.

[5] Y. Deng, B. Manjunath, Unsupervised segmentation of color-texture regions in images and video, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (8) (2001) 800–810.

[6] J. Li, J. Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (9) (2003) 1075–1088.

[7] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories, Computer Vision and Image Understanding 106 (1) (2007) 59–70.

[8] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: Computer Vision–ECCV 2006, Springer, 2006, pp. 1–15.

[9] G. Carneiro, A. B. Chan, P. J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 394–410.

[10] F. Monay, D. Gatica-Perez, Modeling semantic aspects for cross-media image indexing, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1802–1817.

[11] J. Li, J. Z. Wang, Real-time computerized annotation of pictures, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (6) (2008) 985–1002.

[12] X.-J. Wang, L. Zhang, X. Li, W.-Y. Ma, Annotating images by mining image search results, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (11) (2008) 1919–1932.

[13] R. C. Wong, C. H. Leung, Automatic semantic annotation of real-world web images, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (11) (2008) 1933–1944.

[14] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, H.-J. Zhang, Two-dimensional multilabel active learning with an efficient online adaptation model for image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1880–1897.

[15] C. Liu, J. Yuen, A. Torralba, Nonparametric scene parsing via label transfer, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (12) (2011) 2368–2382.

[16] N. Chen, J. Zhu, F. Sun, E. P. Xing, Large-margin predictive latent subspace learning for multiview data analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (12) (2012) 2365–2378.

[17] L. Wu, R. Jin, A. K. Jain, Tag completion for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (3) (2013) 716–727.

[18] Y. Feng, M. Lapata, Automatic caption generation for news images, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (4) (2013) 797–812.

[19] J. Yu, D. Tao, M. Wang, Adaptive hypergraph learning and its application in image classification, Image Processing, IEEE Transactions on 21 (7) (2012) 3262–3272.

[20] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, Image Processing, IEEE Transactions on 23 (5) (2014) 2019–2032.

[21] J. Yu, Y. Rui, B. Chen, Exploiting click constraints and multi-view features for image re-ranking, Multimedia, IEEE Transactions on 16 (1) (2014) 159–168.

[22] J. Yu, Y. Rui, Y. Y. Tang, D. Tao, High-order distance-based multiview stochastic learning in image classification, Cybernetics, IEEE Transactions on 44 (12) (2014) 2431–2442.

[23] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, X. Wu, Visual-textual joint relevance learning for tag-based social image search, Image Processing, IEEE Transactions on 22 (1) (2013) 363–376.

[24] M. Wang, X. Xia, J. Le, X. Zhou, Effective automatic image annotation via integrated discriminative and generative models, Information Sciences 262 (2014) 159–171.

[25] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, T. X. Han, Contextual weighting for vocabulary tree based image retrieval, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 209–216.

[26] M. J. Choi, A. Torralba, A. S. Willsky, A tree-based context model for object recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (2) (2012) 240–252.

[27] M. Marszalek, C. Schmid, Semantic hierarchies for visual object recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2007, pp. 1–7.

[28] J.-R. Kim, H. S. Chang, S. U. Lee, S. Sull, Efficient subtree pruning scheme in tree-structured hierarchy, IEEE Transactions on Circuits and Systems for Video Technology 17 (5) (2007) 635–638.

[29] L. Wang, L. Zhou, C. Shen, L. Liu, H. Liu, A hierarchical word-merging algorithm with class separability measure., IEEE transactions on Pattern Analysis and Machine Intelligence 36 (3) (2014) 417–435.

[30] P. Khanna, S. Pandey, H. Yokota, Finding image semantics from a hierarchical image database based on adaptively combined visual features, in: Database and Expert Systems Applications, Springer, 2013, pp. 103–117.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 248–255.

[32] C. Fellbaum, Wordnet: An electronic lexical database. 1998, WordNet is available from http://www. cogsci. princeton. edu/wn.

[33] X.-Y. Wang, Y.-J. Yu, H.-Y. Yang, An effective image retrieval scheme using color, texture and shape features, Computer Standards & Interfaces 33 (1) (2011) 59–68.

[34] N. S. Vassilieva, Content-based image retrieval methods, Programming and Computer Software 35 (3) (2009) 158–180.

[35] G.-H. Liu, J.-Y. Yang, Content-based image retrieval using color difference histogram, Pattern Recognition 46 (1) (2013) 188–198.

[36] M. A. Stricker, M. Orengo, Similarity of color images, in: IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology, International Society for Optics and Photonics, 1995, pp. 381–392.

[37] S. Sural, G. Qian, S. Pramanik, A histogramm with perceptually smooth color transition for image retrieval., in: JCIS, 2002, pp. 664–667.

[38] X. Wang, Z. Wang, A novel method for image retrieval based on structure elements descriptor, Journal of Visual Communication and Image Representation 24 (1) (2013) 63–74.

[39] B. S. Manjunath, W.-Y. Ma, Texture features for browsing and retrieval of image data, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8) (1996) 837–842.

[40] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[41] Y. Rubner, C. Tomasi, L. J. Guibas, A metric for distributions with applications to image databases, in: Sixth International Conference on Computer Vision, IEEE, 1998, pp. 59–66.

[42] NIST/SEMATECH e-Handbook of Statistical Methods (October 2014).
URL `http://www.itl.nist.gov/div898/handbook/eda/section3/eda356.htm`

[43] Google image search engine (October 2014).
URL `www.google.com`

[44] F. Malik, B. Baharudin, Quantized histogram color features analysis for image retrieval based on median and laplacian filters in dct domain, in: International Conference on Innovation Management and Technology Research (ICIMTR), IEEE, 2012, pp. 624–629.

[45] R. Gali, M. Dewal, R. Anand, Genetic algorithm for content based image retrieval, in: Fourth International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), IEEE, 2012, pp. 243–247.

[46] P. Kinnaree, S. Pattanasethanon, S. Thanaputtiwirot, S. Boontho, Rgb color correlation index for image retrieval, Procedia Engineering 8 (2011) 36–41.



Figure 1: Images with semantics as inferred by human.



Color	Upper	Blue	Blue	Green
	Middle	Grey/White	Grey/White	Green/Blue/Black
	Lower	Green	Green	Green
Semantics		Structure, Housing, Living Quarters	Sports, Golf, Outdoor Game	

Figure 2: Some images with low level color features and high level semantics.

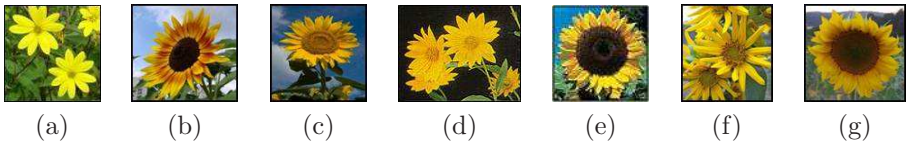


Figure 3: Images from seven categories of *Sunflower* as per ImageNet (a) Swamp (b) Common (c) Giant (d) Showy (e) Maximilian (f) Prairie (g) Jerusalem.

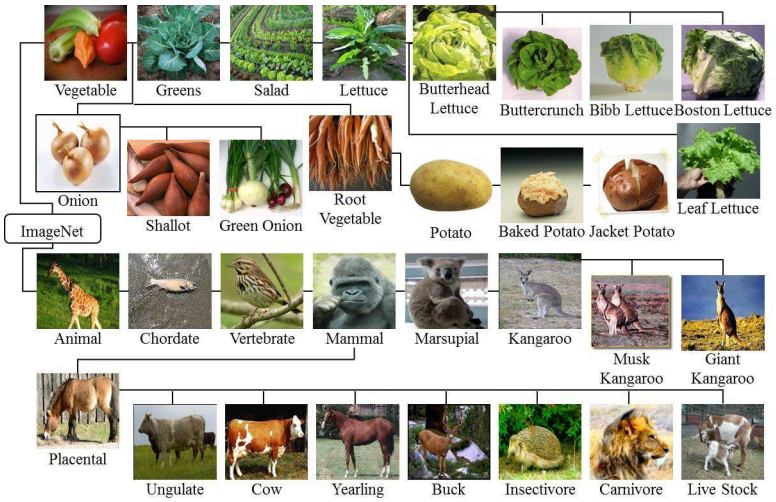


Figure 4: A snapshot of *Vegetable* and *Animal* subtrees of ImageNet 2011 Winter Release.

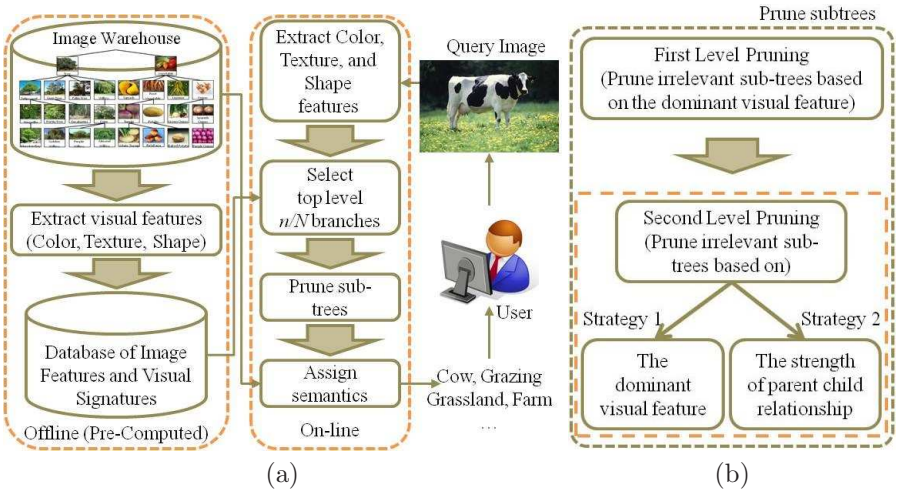


Figure 5: (a) Process flow of the proposed system (b) 'Prune subtrees' module.

Input: A query image and an image tree. Let N is the number of subtrees at any level of hierarchy.
Output: Reduced search space to retrieve semantics of query image. At the most 'n' subtrees are selected at each level for further search ($n \leq N$).

Step 1: Get features of the query image as discussed in Section 6.1.
Step 2: Corresponding to each feature, calculate the distance of query image with roots of each of the N subtrees at the current level of hierarchy.
Step 3: For each feature (color, texture, and shape), select n subtrees ($n \leq N$) having minimum distance from the query image. This results in three lists, one corresponding to each feature, containing n entries.
Step 4: Check these three lists for any of the following three possibilities:

Case 1: Subtree X is the 1st choice in all the lists. It indicates that subtree X has the closest distance with query image with respect to color, texture, and shape features considered. Select only this subtree for further search.

Case 2: Subtree X is the 1st choice in any two lists (representing that X is closer to query image for two features), then select subtree X for further search, and in addition, select (n-1) more subtrees from these two lists having maximum frequency of appearance. Appearance of a subtree in both the lists indicates that this subtree is closer to query image on both the features as compared to other subtrees in the list. In case, subtrees have same frequency, follow *tie breaking criteria*.

Case 3: 1st choice of subtrees in all 3 lists is different. Select top n subtrees based on the maximum frequency of their appearance in these 3 lists. In case of a tie, again follow *tie breaking criteria*.

Step 5: Repeat Step 2 at every level for subtrees selected in Step 4.

Tie breaking criteria:
Having subtrees with the same frequency in the feature lists corresponding to color, texture, and shape features indicate that different subtrees are closer to the query image on these features. In this case select subtrees with minimum sum of weighted distances for color, texture, and shape features.

Figure 6: Flow of execution of the 'Branch Selection Algorithm'.

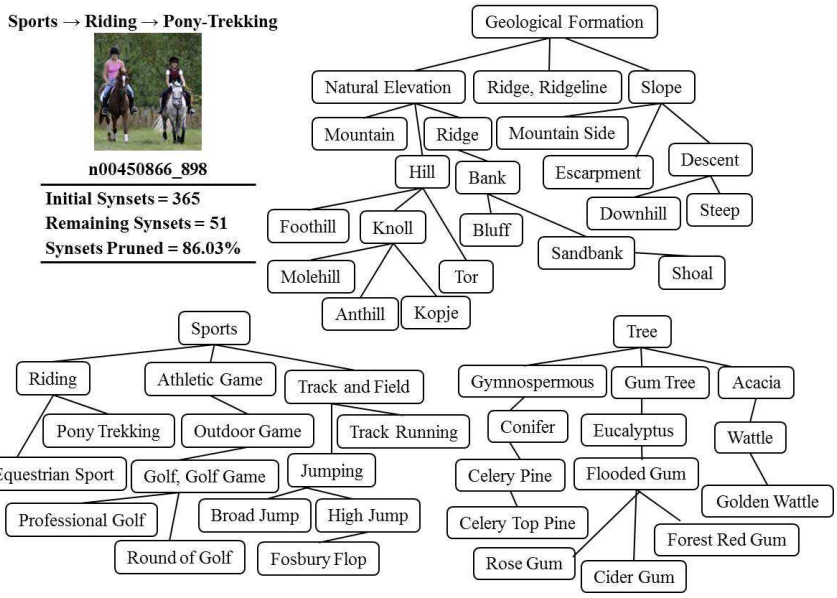


Figure 7: Output of 'Branch Selection Algorithm' (N=11, n=3) for query image "n00450866_898".

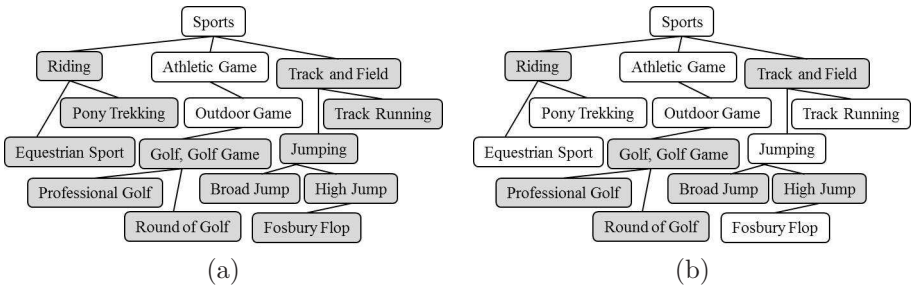


Figure 8: Pruning of Sports subtree selected for the "n00450866_898" (a) Strict Pruning (b) Soft Pruning (gray nodes are the pruned ones).

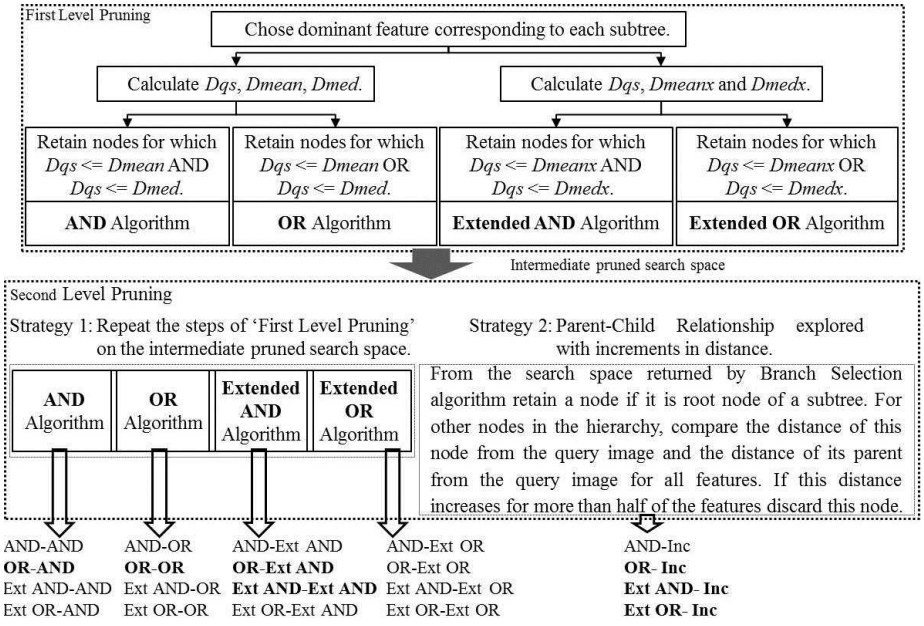
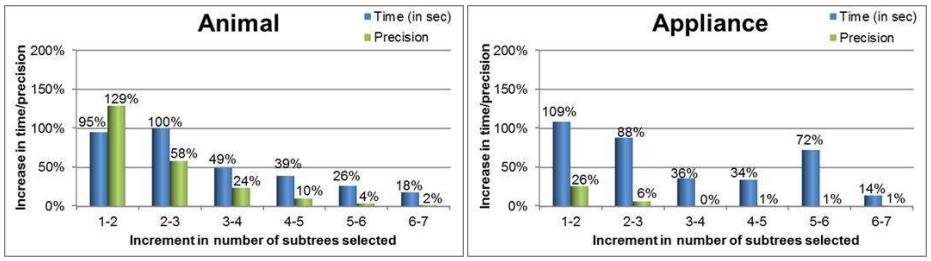


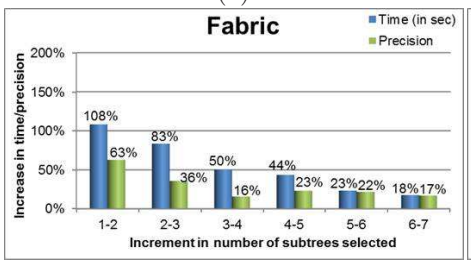
Figure 9: All identified combinations for Pruning Algorithms. ('Ext' and 'Inc' are used to represent Extended and Increasing distances, respectively).


Animal
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
95%
129%
100%
58%
49%
24%
39%
10%
26%
4%
18%
2%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

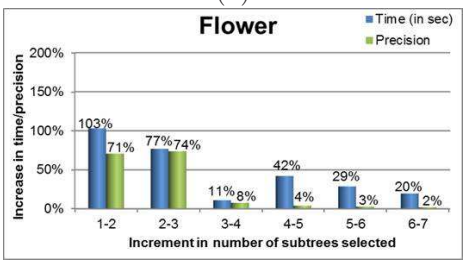
(a)

Appliance
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
109%
26%
88%
6%
36%
0%
34%
1%
72%
1%
14%
1%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

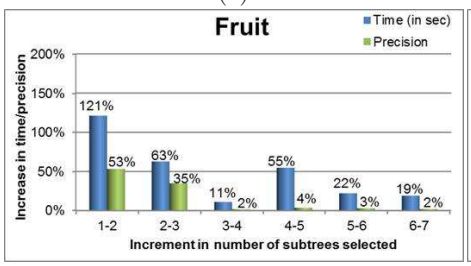
(b)


Fabric
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
108%
63%
83%
36%
50%
16%
44%
23%
23%
22%
18%
17%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

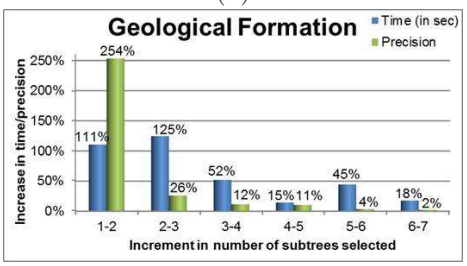
(c)


Flower
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
103%
71%
77%
74%
11%
8%
42%
4%
29%
3%
20%
2%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

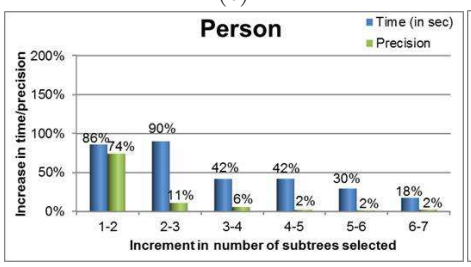
(d)


Fruit
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
121%
53%
63%
35%
11%
2%
55%
4%
22%
3%
19%
2%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

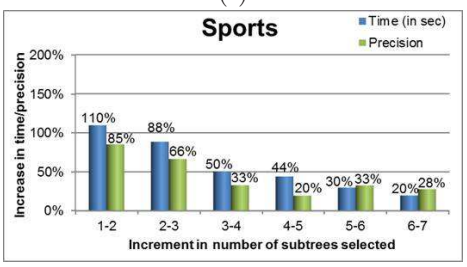
(e)


Geological Formation
Time (in sec)
Precision
Increase in time/precision
250%
200%
150%
100%
50%
0%
111%
254%
125%
26%
52%
12%
15%
11%
45%
4%
18%
2%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

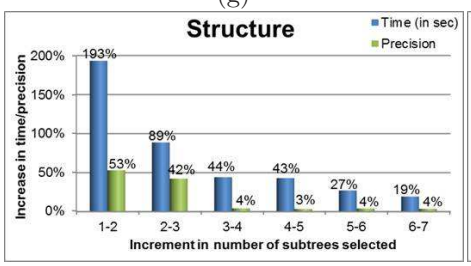
(f)


Person
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
86%
74%
90%
11%
42%
6%
42%
2%
30%
2%
18%
2%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

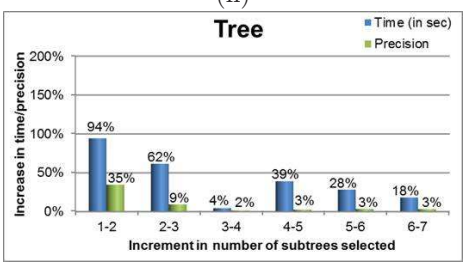
(g)


Sports
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
110%
85%
88%
66%
50%
33%
44%
20%
30%
33%
20%
28%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

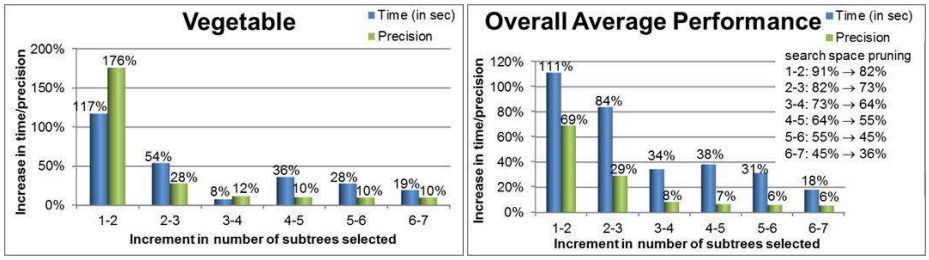
(h)


Structure
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
193%
53%
89%
42%
44%
4%
43%
3%
27%
4%
19%
4%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

(i)

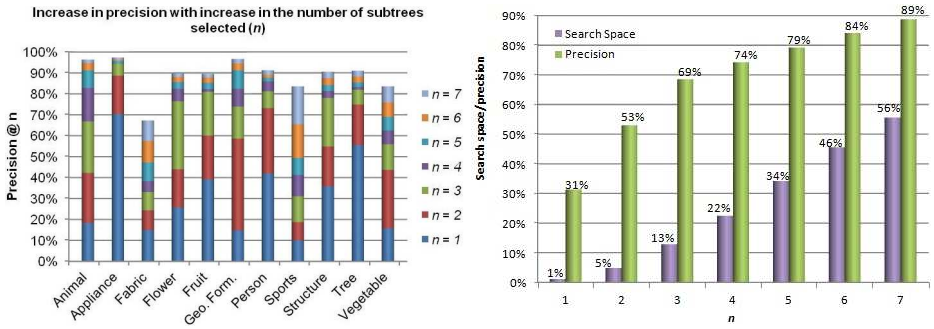

Tree
Time (in sec)
Precision
Increase in time/precision
200%
150%
100%
50%
0%
94%
35%
62%
9%
4%
2%
39%
3%
28%
3%
18%
3%
1-2
2-3
3-4
4-5
5-6
6-7
Increment in number of subtrees selected

(j)



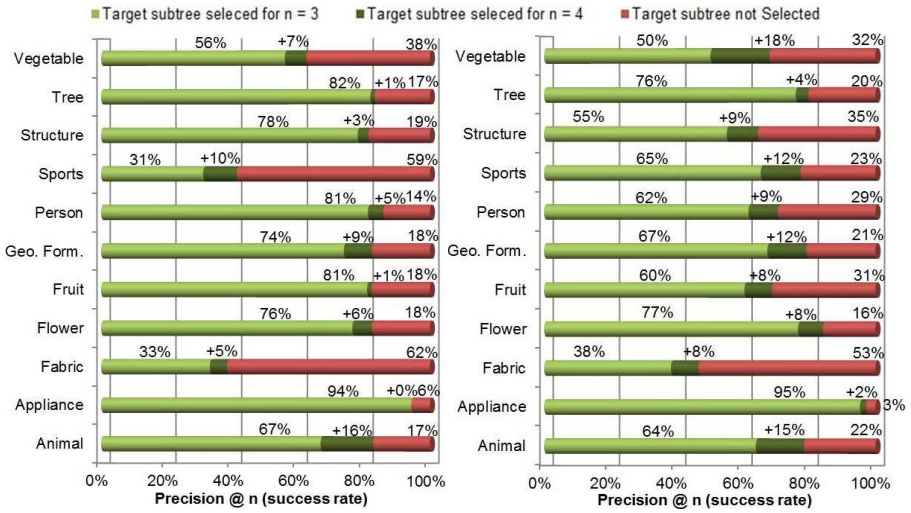
(k) (l)

Figure 10: (a)-(k) Increment in precision and time for set-1 of query images with increasing n for each of the 11 hierarchies; and (l) the same averaged over all hierarchies.



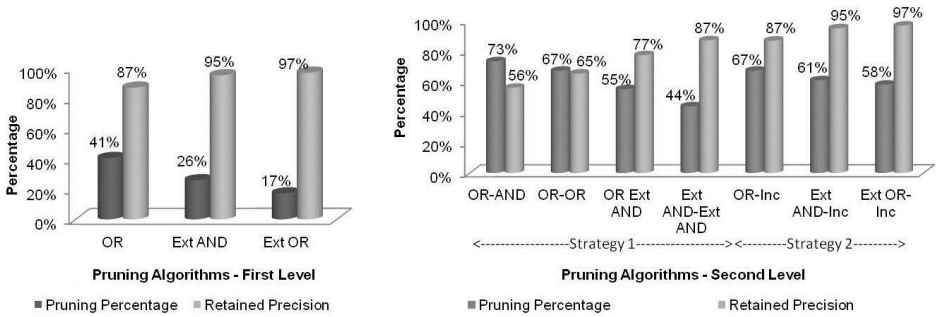
(a) (b)

Figure 11: (a) Increase in precision with increasing value of n for all 11 hierarchies (b) Actual precision and search space used with increasing value of n.



(a) (b)

Figure 12: Performance of 'Branch Selection Algorithm' for (a) set-1 (b) set-2.



(a) (b)

Figure 13: Performance of Pruning Algorithms after (a) First Level (b) Second Level.

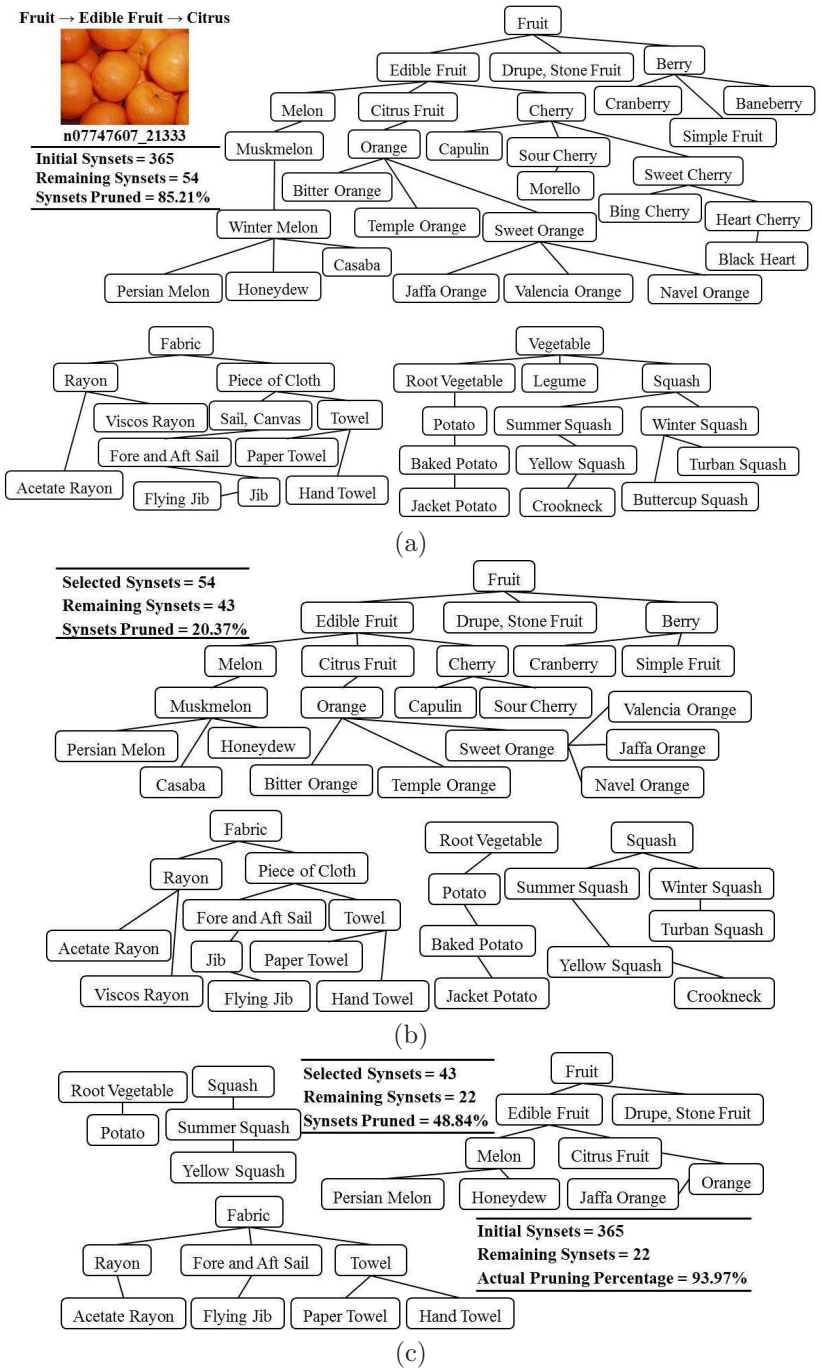


Figure 14: Output for query image "n07747607_21333" (a) Branch Selection Algorithm ($N = 11$, $n = 3$) (b) First Level Pruning (Ext OR) (c) Second Level Pruning (Inc).