# Searching for Images and Videos on the World-Wide Web

*John R. Smith and Shih-Fu Chang*

Department of Electrical Engineering and
Center for Image Technology for New Media,
Columbia University,
New York, N.Y. 10027

{jrsmith, sfchang}@itnm.columbia.edu

Center for Telecommunicatiosn Research
Technical Report #459-96-25

August 19, 1996

**Abstract**

We describe a prototype visual information system for searching for images and videos on the World-Wide Web. New visual information in the form of images, graphics, animations and videos is being published on the Web at an incredible rate. However, cataloging this visual data is beyond the capabilities of current text-based Web search engines. The key to cataloging it is the marriage of text-based processing and content-based visual analysis of the images and videos. In this paper, we describe a complete system by which visual information on the Web is (1) collected by automated agents, (2) processed in both text and visual feature domains, (3) catalogued and (4) indexed for fast search and retrieval. We introduce an image and video search engine which utilizes both text-based navigation and content-based technology for searching visually through the catalogued images and videos. Finally, we provide an initial evaluation based upon the cataloging of over one half million images and videos collected from the Web.

**Keywords** – content-based visual query, image and video storage and retrieval, World-Wide Web.

# 1 Introduction

A large number of catalogs and search engines index the plethora of documents on the World-Wide Web. For example, recent systems, such as Lycos, Alta Vista and Yahoo, index the documents by their textual content. These systems periodically scour the Web, record the text on each page and through processes of automated analysis and/or (semi-) automated classification, condense the Web into compact and searchable indexes. The user, by entering query terms and/or by selecting subjects, uses these search engines to more easily find the desired Web documents. Generally, the text-based Web search engines are evaluated on the basis of the size of the catalog, speed and effectiveness of search and ease of use [1].

However, no tools are currently available for searching for images and videos. This absence is particularly notable given the highly visual and graphical nature of the Web [2]. Visual information is published both as embedded in Web documents and as stand-alone objects. The visual information takes the form of images, graphics, bitmaps, animations and videos. As with Web documents in general, the publication of visual information is highly volatile. New images and videos are added everyday and others are replaced or removed entirely. In order to catalog the visual information, a highly efficient automated system is needed that regularly traverses the Web, detects visual information and processes it in such a way to allow for efficient and effective search and retrieval.

We recently developed a prototype visual information retrieval system[1] to fulfill this need. The system collects images and videos from the Web and provides tools for browsing and searching through the collection. The system is novel in that it utilizes text and visual information synergistically to provide for cataloging and searching for the images and videos. The complete system possesses several powerful functionalities, namely,

- searching using content-based techniques,
- query modification using content-based relevance feedback,
- automated collection of visual information,
- compact presentation of images and videos for displaying query results,
- image and video subject search and navigation,
- text-based searching,
- search results lists manipulations such as intersection, subtraction and concatenation.

We briefly discuss the important concepts in developing our system in the following paragraphs.

## 1.1 Content-based Visual Searching

Content-based technologies have enabled recent advances in the management, search and retrieval of visual information. In particular, content-based techniques provide for the automated assessment of salient visual features such as colors, textures, shapes and spatial information contained within visual scenes. By computing the similarities between images and videos using these extracted visual features, several powerful functionalities are added to the system, which allow

- queries based on the visual features of the data [3],
- automated grouping of scenes into visually homogeneous clusters [4],
- browsing and navigation by content through an image and video archive [5].

[1] `http://www.itnm.columbia.edu/webseek`

### 1.2 Content-based Relevance Feedback

Content-based tools may also be used to improve the query process by learning from the user. In one form of relevance feedback, the user selects an item from the list returned from a query and asks the system to retrieve more that are similar to it in some specified way. The similarity may be based upon visual features, such as colors, textures, spatial layout, or be based upon text or subject classifications. In a second form of relevance feedback [6], the user selects the images and videos from the current results that are most and least typical of the ones desired. From these positive and negative examples, the system automatically reformulates the query to better match the user's instruction. In this sense, the system learns from the user what visual information is desired and converges to it through successive rounds of queries.

### 1.3 Visual Information Collection and Processing

The images and videos are catalogued using a series of automated agents that traverse the Web detecting visual data. The system builds several indexes for the images and videos based upon

- visual features
- terms and key-terms
- assigned subjects
- image/video types.

Image and video subject and type information is derived from both the textual information – such as the Web address and parent document reference text – and visual features. Because the recent taxonomies of knowledge are not well suited for visual information, we developed a new working taxonomy for image and video subject matter. The taxonomy is based upon a graph structure and contains classes for sports, nature, transportation, art and so forth. The images and videos are classified into subjects using several fully- and semi- automated procedures. In the first process, we utilize a key-term dictionary which prescribes subject classes to the images and videos based upon detection of certain key-terms. In the second process, the directory portion of the image and video Web addresses is parsed and analyzed such that groups of images and videos are flagged for inspection and manual classification.

### 1.4 Search List Manipulation

After receiving a list of images and videos in response to a query, the user may manipulate the search results list by adding and removing items using operations with previous or subsequent search results lists. By tracking the user's queries and search results, the system allows for a large variety of ways for the user to refine queries and browse. The system allows the user to concatenate, subtract and intersect search results lists. Search results list manipulation is of special advantage for visual information because the user may also manipulate and search for images and videos using visual features.

### 1.5 Outline

In this paper we describe the complete system for cataloging and searching for images and videos on the Web. In section 2, we describe in detail the process for automated collection of the visual information. In section 3, we describe the procedures for classifying the collected images and videos using key-term mappings and directory names. We also present and utilize a new taxonomy for visual information. In section 4, we describe the system for navigating through subject classes, searching, viewing query results and manipulating the search

results lists. In section 5, we describe several content-based tools for searching, browsing and revising queries. In particular, we describe the system's utilization of color histograms for the content-based manipulation of images and videos. Finally, in section 6, we provide an initial evaluation of the system in the collected of more than one half million images and videos belonging to 16,773 sites on the World-Wide Web.

## 2 Image and Video Collection Process

The image and video collection process is conducted by several autonomous Web agents or spiders. The agents traverse the Web by following the hyperlinks between documents. They detect images and videos, download and process them and add the new information to the catalog. The overall collection process, illustrated in Figure 1, is carried out by several distinct spiders: (1) *Spider 1* – assembles lists of candidate Web pages that may include images, videos or hyperlinks to them, (2) *Spider 2* – extracts the *URL*s of the images and videos, (3) *Spider 3* – retrieves and analyzes the images and videos.

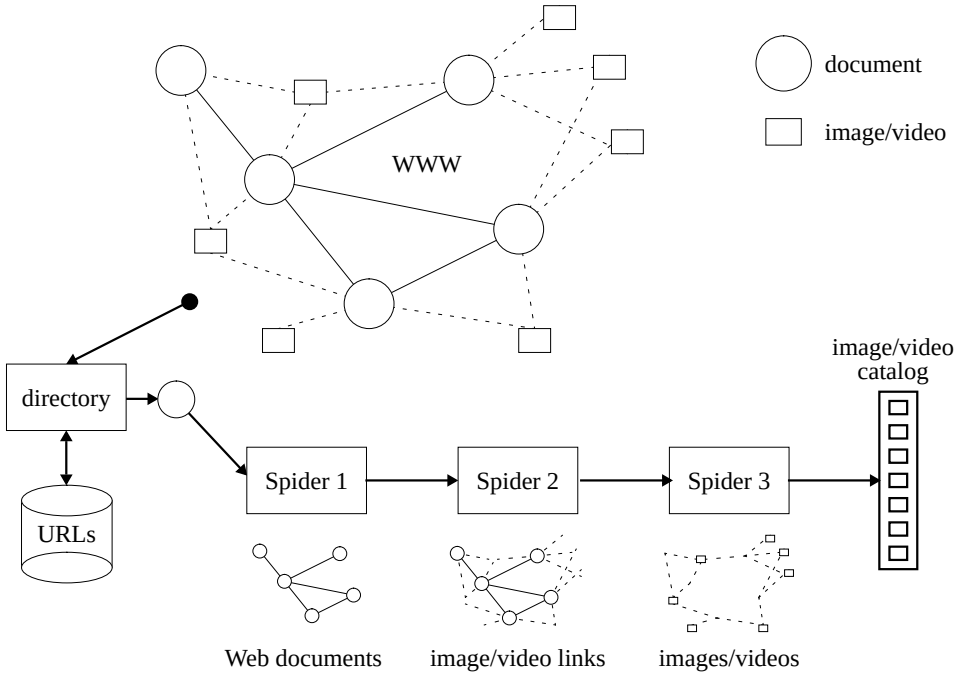



Figure 1: Image and video gathering process via three spiders.

### 2.1 Image and Video Detection

The first phase of the process consists of the two spiders that traverse the Web looking for images and videos, as illustrated in Figure 2. Starting from seed *URL*s, *Spider 1* follows a breadth-first search across the Web. It downloads pages via the Hypertext Transfer Protocol (*HTTP*) protocol and passes the Hypertext Markup Language (*HTML*) code to *Spider 2*. In turn, *Spider 2*, detects new *URL*s, encoded as *HTML* hyperlinks, and adds them back to the queue of Web pages to be downloaded by *Spider 1*. In this sense, *Spider 1* is similar to many of the conventional spiders or robots that follow hyperlinks in some fashion across the Web. [7].

*Spider 2* detects all hyperlinks in the Web documents and converts the relative *URL*s to absolute addresses. By examining the types of the hyperlinks and the filename extensions of the *URL*s, *Spider 2* assigns each *URL* to one of several categories: image, video or *HTML*. The mapping between filename extensions and Web

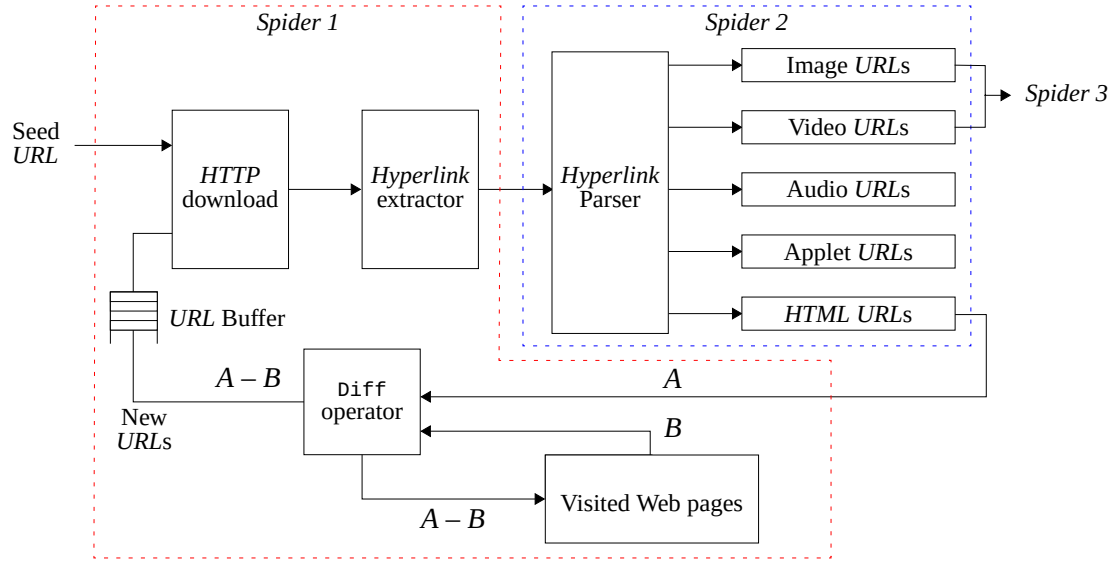



Figure 2: *Spider 1* and *Spider 2* traverse the Web and assemble lists of *URL*s of images and videos.

object type is given by the *Multipurpose Internet Mail Extensions* (MIME) content type labels, as illustrated in Table 1.

| Extension | Type |
|---|---|
| .gif | Compuserve image format |
| .jpg, .jpeg, .jpe, .jfif, .pjpeg, .pjp | JPEG image format |
| .qt, .mov, .moov | Quicktime video format |
| .mpeg, .mpg, .mpr, .mpv, .vbs, .mpegv | MPEG video format |
| .avi | Microsoft video format |
| .htm, .html | Hypertext Markup Language |

Table 1: MIME mapping between extensions and object types.

In the second phase, the list of image and video *URL*s from *Spider 2* is input into *Spider 3*. *Spider 3* retrieves the images and videos, processes them and adds them to the catalog. Three important functions of the *Spider 3* are to

1. extract visual features that allow for content-based techniques in searching, browsing and grouping,

2. extract other attributes such as width, height, number of frames, type of visual data, and so forth,

3. generate an icon, or motion icon, that sufficiently compacts and represents the visual information.

The tasks of *Spider 3* are illustrated in Figure 3. The process of extracting visual features from the images and videos generates color histograms, which are discussed in Section 5. The other attributes of the images and videos populate the database tables, which are defined in Section 3.4. Finally, *Spider 3* generates coarse and highly compressed versions of the images and videos to provide pictorial data in the query output.

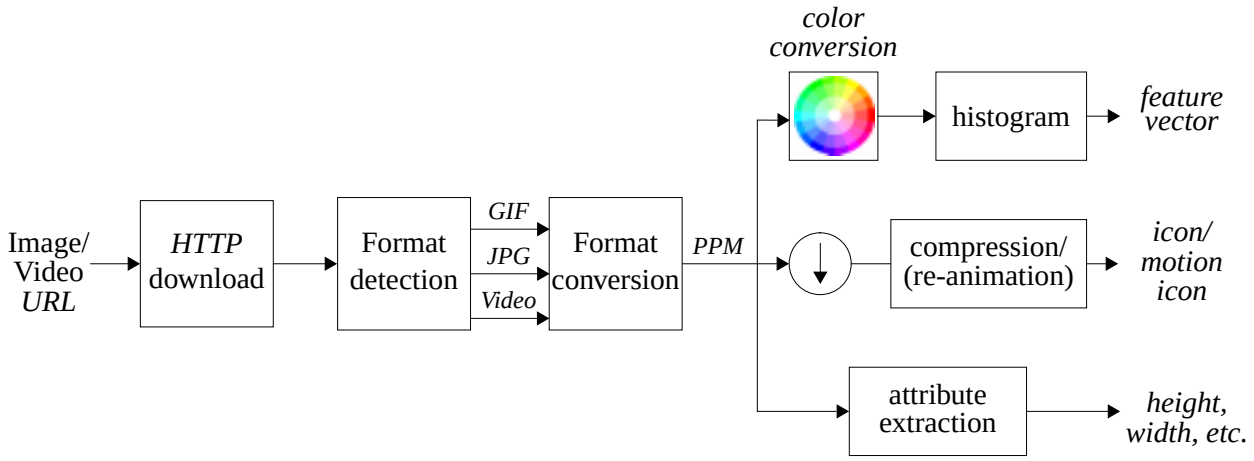



Figure 3: *Spider 3* processes each image/video.

### 2.1.1 Image and Video Presentation

For images, the coarse versions are obtained by simply subsampling and compressing the originals where the compression format, either JPEG or GIF, is chosen to match the original image format. For video, the coarse versions are generated by subsampling the original video both spatially and temporally. The temporal subsampling is achieved in a two step process: first, one frame is kept every one second of video. Next, scene change detection is performed on the frames to detect the key frames of the sequence [8]. This allows for the elimination of duplicate scenes in the coarse version. Finally, the video is re-animated from the key frames and packaged as an animated GIF file. Upon retrieval from a query, the coarse videos appear to the user as animated samples of the original video.

## 3 Subject Classification and Indexing

Utilization of text is essential to the cataloging process. In particular, every image and video on the Web has a unique Web address and possibly other *HTML* tags, which provide for valuable interpretation of the visual information. We process the Web addresses, or *URL*s, and *HTML* tags in the following ways to index the images and videos:

- term extraction
- directory name extraction
- automated key-term to subject mapping using the key-term dictionary and
- semi-automated directory name to subject mapping.

### 3.1 Text Processing

Images and videos are published on the Web in two forms: *inlined* and *referenced*. The *HTML* syntax differs in the two cases. To inline, or embed, an image or video in a Web document, the following code is included in the document: `<img src=URL alt=[alt text]>`, where `URL` gives the relative or absolute address of the image or video. The optional `alt` tag specifies the text that may appear in place of the image or video when the browser is loading the image/video or has trouble finding or displaying the visual data. Alternatively, images and videos

may be referenced from parent Web pages using the following code: `<a href=URL>`[hyperlink text]`</a>`, where the optional [hyperlink text] provides the high-lighted text that describes the object pointed to by the hyperlink, in this case, an image or video.

### 3.1.1 Term Extraction

The **terms** are extracted from the image and video *URL*s, *alt* tags and hyperlink text by chopping the text at non-alpha characters. For example, the *URL* of an image or video has the following form

`URL = http://host.site.domain[:port]/[ user/][directory/][file[.extension]]`

where `[...]` denotes an optional argument. For example, several typical *URL*s are

$\text{URL}_1$ = `http://www.mynet.net:80/animals/domestic-beasts/dog37.jpg`,
$\text{URL}_2$ = `http://camille.gsfc.nasa.gov/rsd/movies2/Shuttle.gif`,
$\text{URL}_3$ = `http://www.arch.columbia.edu/DDL/projects/amiens/slides/slide6b.gif`.

Terms are extracted from the `directory` and `file` strings using $\mathcal{F}_{\text{key}}$ and $\mathcal{F}_{\text{chop}}$ where

$$\mathcal{F}_{\text{key}}(\text{URL}) = \mathcal{F}_{\text{chop}}(\texttt{directory/file}),$$

where $\mathcal{F}_{\text{chop}}(\text{string})$ = list of substrings that are delimited by non-alpha characters. For example,

$$\mathcal{F}_{\text{key}}(\text{URL}_1) = \mathcal{F}_{\text{chop}}(\text{“animals/domestic-beasts1/dog37”}) = \text{“animals”, “domestic”, “beasts”, “dog”}.$$

For one, the terms allow text-based searching via string-matching. After extracting the terms, the system indexes the images and videos directly using inverted files. The process of file-inversion is illustrated in Tables 2. For example, if the user enters the query term “animal”, the images and videos with IMID = 259503 and 106441 are retrieved, respectively. In addition, certain terms, key-terms, are used to map the images and videos to subject classes, as we explain shortly.

| IMID | Terms |
|---|---|
| 121216 | nasa, clipart |
| 259503 | animal, dog |
| 151285 | astronomy, nasa |
| 106441 | animal, clipart |

| Terms | IMID |
|---|---|
| animal | 259503, 106441 |
| astronomy | 151285 |
| clipart | 121216, 106441 |
| dog | 259503 |
| nasa | 121216, 151285 |

Table 2: File inversion of terms for images/videos.

### 3.1.2 Directory Name Extraction

A **directory name** is a phrase extracted from the URLs that groups images and videos by location on the Web. The directory name consists of the directory portion of the *URL*, namely, $\mathcal{F}_{\text{dir}}(\text{URL}) = \texttt{directory}$. For example, $\mathcal{F}_{\text{dir}}(\text{URL}_1)$ = “animals/domestic-beasts”. The directory names are also used by the system to map images and videos to subject classes.

### 3.2 Key-term Dictionary and Directory Name to Subject Mappings

A **key-term** is a manually identified term that corresponds to one or more subject classes. The **key-term dictionary** contains the set of key-terms and their corresponding mappings to subject classes. We build the key-term dictionary in a semi-automated process. In the first stage, the term histogram for the image and video archive is computed. Then the terms are ranked by frequency and are presented for manual assessment. Ranking the terms in order of highest frequency prioritizes them for inspection. The goal of the manual assessment is to determine if a term can be assigned to the key-term dictionary. To make the decision, we consider the descriptive ability of the term and its possible correspondence to one or more subject classes. Terms with multiple meanings make poor key-terms. For example, the term "rock" is a not a good key-term due to its possible disparate references to either stone, or rock music, or several other things. Once a term and its mappings are added to the key-term dictionary, it applies to all existing and new images and videos.

| Non-descriptive | |
|---|---|
| term | count |
| image | 86380 |
| gif | 28580 |
| icon | 14798 |
| pic | 14035 |
| img | 14011 |
| graphic | 10320 |
| picture | 10026 |
| small | 9442 |
| art | 8577 |
| gallery | 6989 |
| thumb | 6669 |

(a)

| Descriptive key-terms and mappings | | |
|---|---|---|
| key-term | count | mapping to subject |
| planet | 1175 | astronomy/planets |
| music | 922 | entertainment/music |
| texture | 831 | graphics/textures |
| aircraft | 458 | transportation/aircraft |
| travel | 344 | travel |
| astronomy | 320 | astronomy |
| gorilla | 273 | animals/gorillas |
| starwars | 204 | entertainment/movies/films/starwars |
| soccer | 195 | sports/soccer |
| dinosaur | 180 | animals/dinosaurs |
| porsche | 139 | transportation/automobiles/porsches |

(b)

Table 3: Sample (a) term counts and (b) key-terms, counts and subject mappings for 500,000 images and videos.

From the initial experiments of cataloging 500,000 images and videos, the terms listed in Table 3 are a sample of those extracted. Notice in Table 3(a) that some of the most common terms are not sufficiently descriptive of the visual information, i.e., terms "image", "picture". However, the terms in Table 3(b) clearly indicate the subject of the images and videos, i.e., terms "aircraft", "gorilla", "porsche". These key-terms are extremely useful for classifying the images and videos into subject classes. For example, we added the key-terms and corresponding subject mappings illustrated in Table 3(b) to the key-term dictionary.

In a similar process, the directory names are inspected and manually mapped to subject classes. Very often an entire directory of images/videos corresponds to a particular topic and can be mapped to one or more subject classes. Similar to the process for key-term identification, the system computes the histogram of directory names and presents it for manual inspection. A directory that sufficiently groups images and videos related to a particular topic is then mapped to the appropriate subject classes.

In Section 6.1, we demonstrate that these methods of key-term and directory name identification and subject mapping provide excellent performance in classifying the images and videos by subject. We also hope that by incorporating some results of natural language processing [9], in addition to using visual features, we can further improve and automate the subject classification process.

### 3.3 Image and Video Subject Taxonomy

A **subject class** or **subject** is an ontological concept that represents the semantic content of an image or video, i.e., "basketball". A **subject taxonomy** is an arrangement of subject classes into an is-a hierarchy. We are developing a new subject taxonomy for image and video subject matter, a portion is illustrated in Figure 4, in the process of inspecting the terms for key-term mappings, as described above. For example, when a new and descriptive term, such as "basketball" is detected and added to the key-term dictionary, we add a corresponding subject class to the taxonomy if it does not already exist, i.e., "sports/basketball".

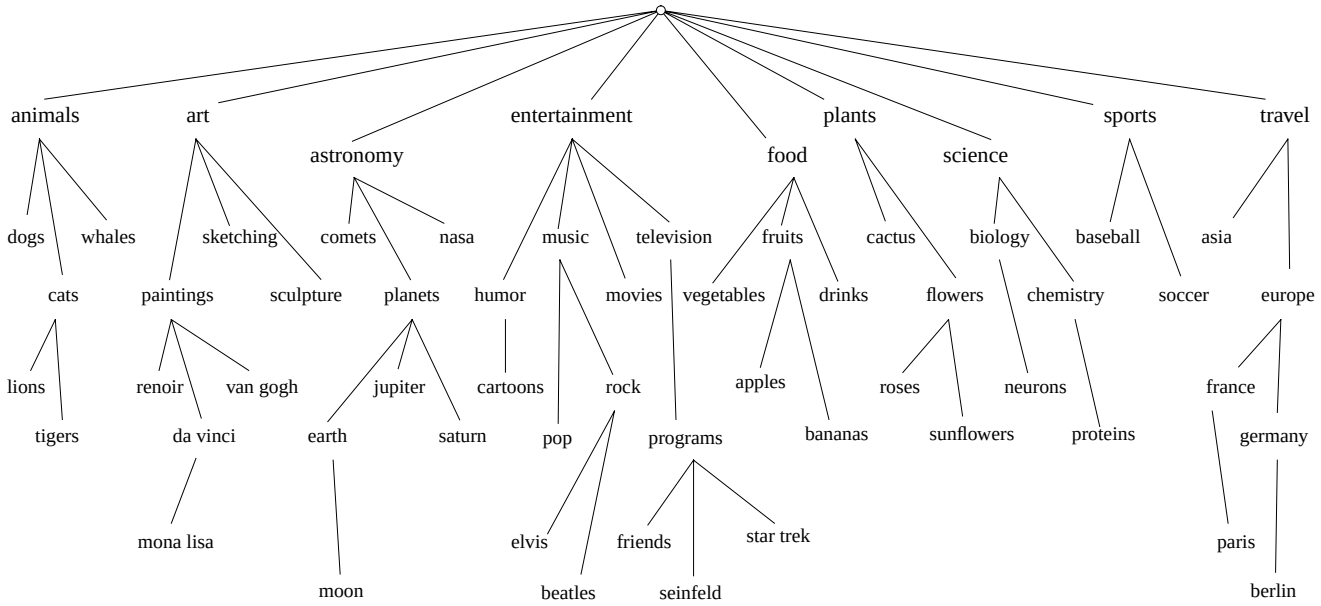



Figure 4: Portion of the image and video subject taxonomy.

### 3.4 Catalog Database

As described above, each retrieved image and video is processed and the following information tables are populated:

| | | |
|---|---|---|
| IMAGES | – | $\overline{\texttt{IMID}}$ URL NAME FORMAT WIDTH HEIGHT FRAME DATE TEXT |
| TYPES | – | $\overline{\texttt{IMID}}$ $\overline{\texttt{TYPE}}$ |
| SUBJECTS | – | $\overline{\texttt{IMID}}$ $\overline{\texttt{SUBJECT}}$ |
| TEXT | – | $\overline{\texttt{IMID}}$ $\overline{\texttt{TERM}}$ |
| FV | – | $\overline{\texttt{IMID}}$ COLOR-HISTOGRAM |

where special (non-alphanumeric) data types are given as follows:

| | | |
|---|---|---|
| TYPE | $\in$ | {Color photo, Color graphic, Video, B/w image, Gray image} |
| SUBJECT | $\in$ | {Subject classes from taxonomy in Figure 4} |
| COLOR-HISTOGRAM | $\in$ | $\mathcal{R}^{166}$ (166-bin histogram). |

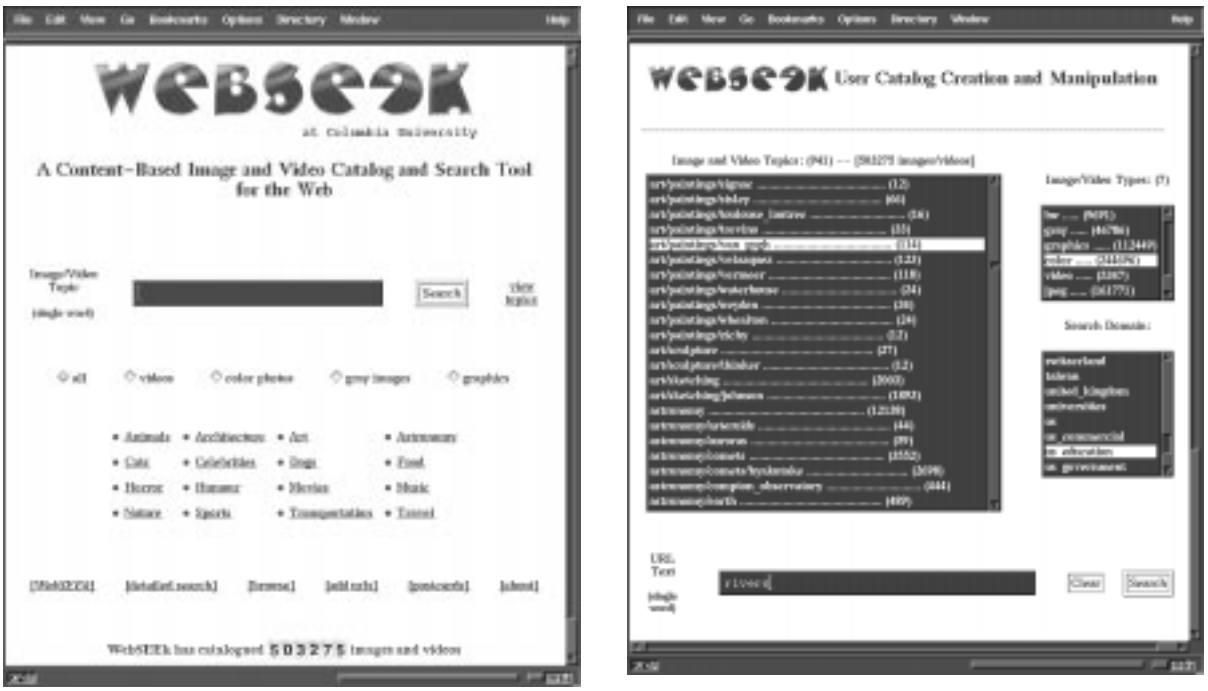
The automated assignment of TYPE to the images and videos using visual features is explained in Section 5.2. Queries on the database tables: IMAGES, TYPES, SUBJECTS and TEXT are performed using standard rela-

tional algebra. For example, the query: `Give me all records with TYPE` = "video", `SUBJECT` = "news" and `TERM` = "basketball" can be carried in SQL as follows:

**SELECT** `IMID`
**FROM** `TYPES, SUBJECTS, TEXT`
**WHERE** `TYPE =` "video" **AND** `SUBJECT =` "news" **AND** `TERM =` "basketball".

However, content-based queries, which involve table `FV`, require special processing, which is discussed in more detail in Sections 4.2 and 5.

# 4 Search, Browse and Retrieval




(a) (b)

Figure 5: Main screens for (a) searching by selecting from several subjects or by text, (b) detailed subject navigation.

To search for images and videos, the user issues a query, which extracts items from the catalog. The user-interface and main search screens are illustrated in Figure 5(a) and (b). The user may initiate the search by entering terms or by selecting subjects directly. The overall search process and model for user-interaction is depicted in Figure 6. As illustrated, a query for images and videos produces a search results list, list $A$, which is presented to the user. For example, Figure 7(a) illustrates the search results list for a query for images and videos related to "nature", that is, $A$ = Query(`SUBJECT` = "nature"). The user may manipulate, search or view $A$.

## 4.1 Search Results List Manipulation

After possibly searching and/or viewing the search results list, it is fed-back to the manipulation module as list $C$, as illustrated in Figure 6. The user manipulates $C$ by adding or removing records. This is done by issuing

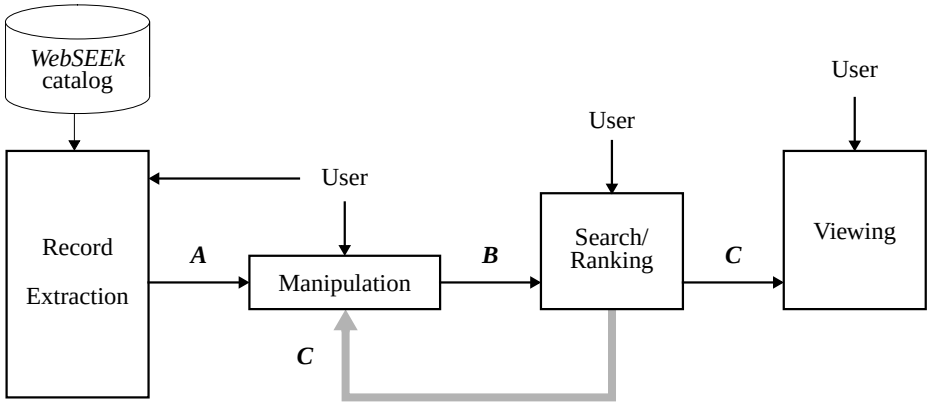



Figure 6: Search, retrieval and search results list manipulation processes.

a new query that creates a second, intermediate list $A$. The user then generates a new search results list $B$ by selecting one of the following manipulations on $C$ using $A$, for example, define $C$ = Query(`SUBJECT` = "nature") and $A$ = Query(`TERM` = "sunset"), then

| | | |
|---|---|---|
| union: | $B = A \cup C$, | i.e., $B$ = Query(`TERM` = "sunset" or `SUBJECT` = "nature"), |
| intersection: | $B = A \cap C$, | i.e., $B$ = Query(`TERM` = "sunset" and `SUBJECT` = "nature"), |
| subtraction: | $B = C - A$, | i.e., $B$ = Query(`SUBJECT` = "nature" and `TERM` $\neq$ "sunset"), |
| replacement: | $B = A$, | i.e., $B$ = Query(`TERM` = "sunset"), |
| keep: | $B = C$, | i.e., $B$ = Query(`SUBJECT` = "nature"). |

### 4.2 Content-based visual query

The user may browse and search the list $B$ using both content-based and text-based tools. In the case of content-based searching, the output is list $C$, where $C \subseteq B$ gives an ordered subset of $B$, and $C$ is ordered by highest similarity to the user's selected item. In the current system, list $C$ is truncated to $\mathcal{N} = 60$ records, where $\mathcal{N}$ may be adjusted by the user. The search and browse operations may be conducted on the input list $B$ or on the entire catalog at any time. In the first case, the user may browse the current search results list by selecting one of the items and instructing the system to reorder the list by highest similarity to the selected item.

For example, $C = B \simeq B^{\text{sel}}$, where $\simeq$ means visual similarity, ranks list $B$ in order of highest similarity to a selected item from $B$. For example, the following content-based visual query:

$$C = \text{Query}(\texttt{SUBJECT} = \text{"nature"}) \simeq B^{\text{sel}}(\text{"mountain scene image"}),$$

ranks the "nature" images and videos in order of highest visual similarity to the selected "mountain scene image." Alternatively, the user can select one of the items in the current search results list $B$ to search the entire catalog for similar items. For example, $C = A \simeq B^{\text{sel}}$ ranks list $A$, where $A$ is the full catalog, in order of highest visual similarity to the selected item from $B$. In the example illustrated in Figure 7(b), the query $C = A \simeq B^{\text{sel}}$("red race car"), retrieves the images and videos from the full catalog that are most similar to the selected image of a "red race car."


(a) (b)

Figure 7: (a) Search results for `SUBJECT` = "nature", (b) content-based visual query results for images/videos $\simeq$ "red race car".

### 4.3 Search Result List Views

The user has several options for viewing search results. Since the visual information requires more communication bandwidth than text, the user is given control in viewing and browsing the search results to enable them to be inspected quickly. The default view presents for each catalog record a small (approximately $96 \times 96$ pixels) icon for each image and video scene in addition to other relevant fields, see Figure 7(a) and (b). Alternatively, the user can select to eliminate the display of the icon altogether, in which case the name of the image/video is displayed. Only $\mathcal{L} = 15$ records at a time are presented to the user in the view. The system gives the user controls to navigate the list by getting the *next*, *previous* and *top* $\mathcal{L}$ records. The user may conveniently select an item for full download, which retrieves the image/video from the original *URL* to the user.

## 5 Content-based Techniques

The system provides tools for content-based searching for images and videos using color histograms generated from the visual scenes. Recent research has explored several types of features for content-based visual query. Certain feature sets are suited to particular application domains, for example, the management of satellite or medical imagery. We adopted color histograms in the prototype system in order to utilize a domain-independent approach. The content-based techniques developed here for indexing, searching and navigation can be applied, in principle, to other types of features and application domains.

The color histograms describe the distribution of colors in each image or video. We define the color histograms as discrete, 166 bin, distributions in a quantized $HSV$ color space [5]. The system computes a color histogram for each image and video scene, which is used to assess its similarity to other images and video scenes. The color histograms are also used to automatically assign the images and videos to type classes using

Fisher discriminant analysis, as described in Section 5.2.

### 5.1 Color Histograms Similarity

The histogram dissimilarity function measures the weighted dissimilarity between histograms. For example, the quadratic distance between query histogram $\mathbf{h}_q$ and target histogram $\mathbf{h}_t$ is given by:

$$d_{q,t} = (\mathbf{h}_q - \mathbf{h}_t)^t \mathbf{A} (\mathbf{h}_q - \mathbf{h}_t), \tag{1}$$

where $\mathbf{A} = [a_{i,j}]$ is a symmetric matrix and $a_{i,j}$ denotes the similarity between colors with indexes $i$ and $j$ such that $a_{i,i} = 1$. Note that the histograms are normalized such that $||\mathbf{h}|| = 1$, where $||\mathbf{h}|| = \sqrt{\sum_{m=0}^{M-1} h[m]^2}$.

In order to achieve high efficiency in the color histogram query process, we decompose the color histogram quadratic formula. This provides for efficient computation and indexing. By defining $\mu_q = \mathbf{h}_q^t \mathbf{A} \mathbf{h}_q$, $\mu_t = \mathbf{h}_t^t \mathbf{A} \mathbf{h}_t$ and $\mathbf{r}_t = \mathbf{A}\mathbf{h}_t$, the color histogram quadratic distance is given as

$$d_{q,t} = \mu_q + \mu_t - 2\mathbf{h}_q^t \mathbf{r}_t. \tag{2}$$

By partitioning vector $\mathbf{r_t}$ into elements $r_t[m]$'s, the distance function can be approximated to arbitrary precision by setting $\tau$ in

$$d_{q,t} - \mu_q = \mu_t - 2 \sum_{\forall m \text{ where } h_q[m] \geq \tau} h_q[m] r_t[m]. \tag{3}$$

That is, any query for the most similar color histogram to $\mathbf{h}_q$ may be easily processed by storing and indexing individually $\mu_t$ and $r_t[m]$'s, where $m \in 1 \ldots M$. Notice also that $\mu_q$ is a constant of the query. The closest color histogram $\mathbf{h}_t$ is given as the one that minimizes $\mu_t - 2\sum_{\forall m \text{ where } h_q[m] \geq \tau} h_q[m] r_t[m]$. By using the efficient computation described in Eq. 3, we are able to greatly reduce the query processing time, as demonstrated in Section 6.3.

### 5.2 Automated Type Assessment

By training on samples of the color histograms of images and videos, we developed a process of automated type assessment using Fisher discriminant analysis. Fisher discriminant analysis constructs a series of uncorrelated linear weightings of the color histograms that provide for maximum separation between training classes. In particular, the linear weightings are derived from the eigenvectors of the matrix given by the ratio of the between-class to within-class sum-of-square matrices for $K$ classes [10]. New color histograms, $\mathbf{h}_n$ are then automatically assigned to nearest type class $k$ where

$$[\mathbf{T}(\mathbf{h}_n - \mathbf{m}_k)]^2 \leq [\mathbf{T}(\mathbf{h}_n - \mathbf{m}_i)]^2, \quad \forall i \neq k, \tag{4}$$

and where $\mathbf{T}$ is the matrix of eigenvectors derived from the training classes and color histograms, and $\mathbf{m}_i$ is the mean histogram for class $i$. In Section 6.2, we show that this approach provides excellent automated classification of the images and videos into several broad type classes. We hope to further increase the number of type classes and improve the classification performance by incorporating other visual features into the process.

### 5.3 Relevance Feedback

The user can best determine from the results of a query which images and videos are relevant and not relevant. The system can use this information to reformulate the query to better retrieve the images and videos the user desires [6]. Using the color histograms, relevance feedback is accomplished as follows: let $I_r$ = {relevant

images/videos} and $I_n$ = {non-relevant images/videos} as determined by the user. The new query vector $\mathbf{h}_q^{k+1}$ at round $k+1$ is generated by

$$\mathbf{h}_q^{k+1} = ||\alpha \mathbf{h}_q^k + \beta \sum_{i \in I_r} \mathbf{h}_i - \gamma \sum_{j \in I_n} \mathbf{h}_j||, \tag{5}$$

where $||\cdot||$ indicates normalization. The new images and videos are retrieved using $\mathbf{h}_q^{k+1}$ and the distance metric in Eq. 3. One formulation of relevance feedback assigns the values $\alpha = 0$, and $\beta = \gamma = 1$, which weights the positive and negative examples equally. The process of selecting the example images for content-based relevance feedback searching is illustrated in Figure 8(a). A simpler form of relevance feedback allows the user to select only one positive example in order to iterate the query process. In this case, $\alpha = \gamma = 0$, $\beta = 1$, $|I_r| = 1$ and $|I_{nr}| = 0$ gives the new query vector directly from the selected image/video's color histogram, $\mathbf{h}_{I_r}$ as follows,

$$\mathbf{h}_q^{k+1} = \mathbf{h}_{I_r}. \tag{6}$$

### 5.4 Histogram Manipulation

The system also provides a tool for the user to directly manipulate the image and video color histograms to formulate the search. Using the histogram manipulation tool, illustrated in Figure 8(b), the user may select one of the images or videos from the results and display its histogram. The user can then modify the histogram by adding or removing colors. The modified histogram is then used to conduct the next search. The new query histogram $\mathbf{h}_q^{k+1}$ is generated from a selected histogram $\mathbf{h}_s$ by adding or removing colors, which are denoted in the modifications histogram $\mathbf{h}_m$

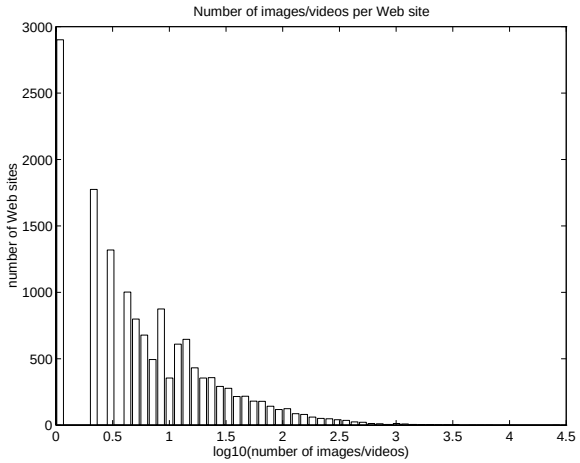
$$\mathbf{h}_q^{k+1} = ||\mathbf{h}_s + \mathbf{h}_m||. \tag{7}$$


(a) (b)

Figure 8: (a) Relevance feedback search allows user to select both positive and negative examples, (b) histogram manipulation allows user to add and remove colors and adjust the color distribution for the next query.

# 6 Evaluation

In the initial trials, the system has catalogued 513,323 images and videos from 46,551 directories on 16,773 distinct Web sites. The process required several months, which was performed simultaneously with the development of the user application. Various information about the catalog process is summarized in Table 4. In all the system has catalogued over 129 Gigabytes of visual information. The local storage of information, which includes coarse versions of the data and color histogram feature vectors, requires approximately 2 Gigabytes.




(a)

| | |
|---|---|
| number of images/videos catalogued | 513,323 |
| number of Web sites providing the images and videos | 16,773 |
| number of distinct Web directories | 46,551 |
| % of catalog is black & white/gray-scale | 14.15% |
| % of catalog is videos | 1.05% |
| % of images and videos classified into subject classes | 68.23% |
| size of subject taxonomy (# classes) | 941 |
| size of key-term dictionary (# terms) | 932 |
| number of directory name to subject mappings | 1018 |

(b)

Table 4: Cataloging of 513,323 images and videos from the Web, (a) number of images/videos per Web site, (b) results.

## 6.1 Subject Classification Evaluation

As indicated in Table 4, the catalog process assigned 68.23% of the images and videos into subject classes using automated mapping for key-terms and semi-automated mapping for directory names. We assessed the subject classification rates for several classes, which is summarized in Table 5(a). The overall performance is excellent, $\sim 92\%$ classification precision. For this assessment, as illustrated in Table 5(a), we chose the classes at random from the subject taxonomy of 941 classes. We established the ground-truth by manually verifying the subject of each image and video in the test sets.

We observed that errors in classification result from several occurrences: (1) key-terms being used out of context by the publishers of the images or videos, (2) the system's reliance on some key-terms that have multiple meanings and contexts, i.e., "madonna" and (3) the system's reliance on key-terms extracted from directory names. For example, in Table 5(a), the precision of subject class "animals/possums" is low because five out of the nine items are not images or videos of possums. These items were classified incorrectly because the key-term "possum" appeared in the directory name. While some of the images in that directory depict possums, others depict only the forests to which the possum are indigenous. When viewed outside of the context of the "possum" web site, the images of forests should not be assigned to the class "animals/possums."

## 6.2 Type Classification Evaluation

We assessed the precision of the automated type classification system, which is summarized in Table 5(b). For this evaluation, both the Training and Test samples consisted of 200 images from each type class. We found the automated type assessment for these five simple classes is quite satisfactory, overall $\sim 95\%$ rate of successful classification. In future work, we will try to extend this system to include a larger number of classes, including new type classes, such as *Fractal* images, *Cartoons*, *Faces*, *Art paintings* and subject classes.

| Subject | # sites | Count | Rate |
|---|---|---|---|
| art/illustrations | 29 | 1410 | 0.978 |
| entertainment/humour/cartoons/daffyduck | 14 | 23 | 1.000 |
| animals/possums | 2 | 9 | 0.444 |
| science/chemistry/proteins | 7 | 40 | 1.000 |
| nature/weather/snow/frosty | 9 | 13 | 1.000 |
| food | 403 | 2460 | 0.833 |
| art/paintings/pissarro | 3 | 54 | 1.000 |
| entertainment/music/mtv | 15 | 87 | 0.989 |
| horror | 366 | 2454 | 0.968 |

(a)

| Type | Rate |
|---|---|
| Color photo | 0.914 |
| Color graphic | 0.923 |
| Gray image | 0.967 |
| B/w image | 1.000 |

(b)

Table 5: Rates of correct (a) Subject classification (precision) for random set of classes and (b) automated type classification.

### 6.3 Efficiency

Another important factor in the image and video search system is the speed at which user operations and queries are performed. In particular, as the archive grows it is imperative that queries do not take so long that they inhibit the user from effectively using the system. In the initial system, the overall efficiency of various database manipulation operations is excellent, even on the large catalog, see Table 6 (server platform = SGI Onyx). In particular, the good performance of the content-based visual query tools is given by the strategies of indexing the 166 bin color histograms described in Section 5.1. For example, the system identifies the $\mathcal{N} = 60$ most similar visual scenes in the catalog of $513,323$ images and videos to a selected query scene in only 1.83 seconds.

| Operation | Time (secs) |
|---|---|
| Direct subject selection | 0.09 |
| Subject class name query | 0.18 |
| Text query | 0.30 |
| Search results lists manipulation | 0.19 to 1.29 |
| Color histogram query (catalog-wide) | 1.83 |
| Color histogram query (list specific) | 1.21 |
| Relevance Feedback query | 1.93 |
| One-way communication of query results and display | 5 to 15 |

Table 6: Execution times of various search, browse and query operations.

## 7 Summary and Future Work

We introduced a new robust system that provides the essential function of cataloging the visual information on the Web. The system automatically collects the images and videos and catalogs them using both textual and visual information. We developed a web application that is very easy to use and provides great flexibility and functionality for browsing and searching for images and videos. In the initial implementation, the system has catalogued and provides searching through more than one half million images and videos.

In future work, we will utilize additional visual features, such as texture, shape and spatial layout, to further enhance the content-based components of the system. In particular, we are porting the *VisualSEEk* [11] system for joint feature/spatial querying to this application. We are also incorporating automated techniques for detecting faces [12] and text in images and videos.

We will also investigate new techniques for exploiting text and visual features independently and jointly to improve the process of cataloging the images and videos and automatically mapping them into subject and type classes. For example, better utilization of the text information in the parent Web pages may provide more information about the images/videos [9]. In addition, several recent approaches for learning from visual features are promising for detecting homogeneities within subject classes and improving the automated classification system. Finally, we will further expand and define the image and video subject taxonomy.

## 8 Acknowledgments

The authors would like to thank Kazi Zaman, Dragomir Radev, Prof. Al Aho and Prof. Kathleen McKeown for their valuable input on this project.

## References

[1] G. S. Jung and V. N. Gudivada. Autonomous tools for information discovery in the world-wide web. Technical Report CS-95-01, School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, 1995.

[2] S. Sclaroff. World wide web image search engines. In *NSF Workshop on Visual Information Management*, Cambridge, MA, June 1995.

[3] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, and C. Faloutsos. The QBIC project: Querying images by content using color, texture, and shape. In *Storage and Retrieval for Image and Video Databases*, volume SPIE Vol. 1908, February 1993.

[4] D. Zhong, H. J. Zhang, and S.-F. Chang. Clustering methods for video browsing and annotation. In *Symposium on Electronic Imaging: Science and Technology – Storage & Retrieval for Image and Video Databases IV*, volume 2670, San Jose, CA, February 1996. IS&T/SPIE.

[5] J. R. Smith and S.-F. Chang. Tools and techniques for color image retrieval. In *Symposium on Electronic Imaging: Science and Technology – Storage & Retrieval for Image and Video Databases IV*, volume 2670, San Jose, CA, February 1996. IS&T/SPIE.

[6] J. J. Rocchio Jr. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 14, pages 313 – 323. Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[7] M. Koster. Robots in the web: threat or treat? *ConneXions*, 9(4), April 1995.

[8] J. Meng, Y. Juan, and Shih-Fu Chang. Scene change detection in a MPEG compressed video sequence. In *Proceedings of the IS&T/SPIE Symposium on Electronic Imaging: Science & Technology*, San Jose, CA, February 1995.

[9] E. J. Guglielmo and N. C. Rowe. Natural-language retrieval of images based on descriptive captions. In *ACM Trans. Info. Systems*, volume 14, pages 237 – 267, July 1996.

[10] K. Fukunaga. *Introduction to Statistical Pattern Recogntion.* Harcourt Brace Javanovich, 1990.

[11] J. R. Smith and S.-F. Chang. Querying by color regions using the *VisualSEEk* content-based visual query system. In M. T. Maybury, editor, *Intelligent Multimedia Information Retrieval.* IJCAI, 1996.

[12] H. Wang and S.-F. Chang. Automatic face region detection in mpeg video sequences. In *Electronic Imaging and Multimedia Systems, SPIE Photonics China '96*, November 1996.