Functionality-Based Web Image Categorization *

Jianying Hu
Avaya Labs Research
233 Mount Airy Road
Basking Ridge, NJ 07920
jianhu@avaya.com

Amit Bagga
Avaya Labs Research
233 Mount Airy Road
Basking Ridge, NJ 07920
bagga@avaya.com

ABSTRACT

The World Wide Web provides an increasingly powerful and popular publication mechanism. Web documents often contain a large number of images serving various different purposes. Identifying the functional categories of these images has important applications including information extraction, web mining, web page summarization and mobile access. This paper describes a study on the functional categorization of Web images using data collected from news web sites. We describe the image categories found in such web pages and their distributions, identify the main research issues involved in automatically classifying images into these categories, and present a novel algorithm for automatic identification of two of the most important image categories, namely story and preview images.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; D.2 [**Software**]: Software Engineering

Keywords

Web document analysis, image classification, text categorization, natural language processing

1. INTRODUCTION

The World Wide Web as an on-line publication mechanism has become increasingly multimedia. Many web documents contain a large number of images, and these images tend to be highly heterogeneous in terms of their functionalities. For example, a news web page may contain images corresponding to specific news stories, icons (*e.g.*, an image representing a sunny forecast), logos, ads, and images containing mostly text serving as section headings, etc.

The automatic identification of the functional categories of images contained in web documents is desirable in many web based information processing tasks, including information extraction, web mining, web page summarization and mobile access. For example, web page filtering and reformatting for wireless access has become a very active research area lately [2, 16, 4, 13]. Proper categorization of images in this case could help prioritize them for transmission over limited bandwidth (presumably, a news image would have higher priority than an ad). Identifying the heading images followed by further processing (*e.g.*, OCR) to recover the textual content could be crucial for both summarization and information extraction. Functional categorization would also effectively reduce the search space for web mining operations including image searching on the web.

While there have been some research activities on the analysis of web images, they have been largely focused on two particular aspects. One is the extraction and recognition of text contained in web images [12, 1]. The other is image search and retrieval on the web [6, 15, 19]. There has been no previous study on functionality based image categorization. Furthermore, while there have been some past analyses on the statistics of images containing text on the web [12, 1, 9], no statistics has been collected on the distributions of other image categories.

In this paper we address the issue of functional categorization of images on the web. For our initial study, we used data collected from news web sites as they generally involve the most diverse set of images and thus pose the greatest challenge for automatic categorization. We describe the image categories found in such web pages and their distributions and point out some research issues. Finally, as a first step towards complete automatic image categorization, we present an algorithm to automatically identify two of the image categories, stories and previews, that are likely of more interest for mobile access. The algorithm makes use of both visual image features and information contained in the surrounding text to separate story and preview images from ads, icons, formatting images, etc. We report experimental results which demonstrate the effectiveness of this algorithm.

2. IMAGE CATEGORIES IN NEWS WEB SITES

The data used in this study was a collection of front pages corresponding to 25 randomly selected news web sites, as listed in Table 1.

Since there has been no previous study on the identification and classification of images found on web pages, we had to define both the categories and their descriptions. After an initial preliminary analysis of the data, we decided to define categories which were functional as opposed to semantic. In other words, an image would be classified into a category based upon its usage in the web page rather than its specific content. This type of image classification has not been studied before and is particularly useful for web summarization and information extraction tasks. The preliminary analysis yielded eight functional categories for the images. Each category is described in detail below with examples shown in Figures 1 and 2. Figure 1 shows the sample images while Figure 2 displays the images in their respective surrounding contexts, which are important in determining their categories.

*(Produces the WWW2003-specific release, location and copyright information). For use with www2003-submission.cls V1.0. Supported by ACM.

WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM xxx.

Table 1: List of web sites.

Description	http address	Description	http address
Arizona Home Page	www.azcentral.com	Denver Post	www.denverpost.com
BBC	www.bbc.co.uk	Detroit Free Press	www.freep.com
BBC News	news.bbc.co.uk	Economist	www.economist.com
BBC Sport	news.bbc.co.uk/sport	Guardian Unlimited	www.guardian.co.uk
Boston Globe	www.boston.com/globe	Houston Chronicle	www.chron.com
Boston Home Page	www.boston.com	LA Times	www.latimes.com
Chicago Tribune	www.chicagotribune.com	Miami Home Page	www.miami.com
CNN	www.cnn.com	New York Times	www.nytimes.com
CNN Sport	sportsillustrated.cnn.com	SF Chronicle	www.sfgate.com/chronicle
CNN Money	money.cnn.com	Sun Spot	www.sunspot.net
Canada Home Page	canada.com/national	Telegraph	www.telegraph.co.uk
Canoe Home Page	canoe.ca	USA Today	usatoday.com
Dallas Morning News	www.dallasnews.com		

Story Images (S) This category contains images whose content is associated with a story that appears on the page. It should be noted that the story associated with the image need not necessarily be present in its entirety on the page. The most common examples of such images are those associated with the top news stories of the day. Figure 2(a) contains an example image that is associated with the story on a game in the recently concluded soccer world cup: "Senegal Stuns France."

Preview Images (P) Images belonging to this category are those whose content is associated with a preview to a story. In other words, the text associated with the image refers to (is a preview of) an actual story appearing elsewhere in the news site. For example, Figure 2(b) contains an image which previews an upcoming movie, presumably discussed in the page pointed to by the hyperlink below the image. The actual text corresponding to the image is: "Summer movie preview: Get a sneak peek at this season's hot flicks ."

Commercial Images (C) These images act as advertisements. Figure 1(c) shows an example of such an image. As seen from Figure 2(c), commercial images are often inserted randomly among unrelated text items.

Host Images (A) These are images of hosts of regular columns or programs, often used to represent the column/program on the front page (for example: Larry King's image that represents his show at the CNN website). An example is shown in Figure 1(d)

Heading Images (H) These images usually contain mainly text and serve as headings for columns and sections. Figure 2(e) contains an example of such an image (as shown in Figure 1(e)).

Icons and Logos (I) This category consists of images that represent specific concepts such as company logos (for example, logo for New York Times), and other commonly used icons such as the image of the sun to represent a sunny weather forecast. An example is shown in Figure 1(f).

Formatting Images (F) These images are used for formatting purposes. Some examples are: horizontal line, vertical line, etc.

Miscellaneous Images (M) This category consists of images that cannot be classified into one of the other categories.

Table 2: Distribution of the images.

Category	# of Images	Percentage
Story (S)	91	10.1
Preview (P)	16	1.8
Host (A)	9	1.0
Commercial (C)	110	12.2
Icons and Logos (I)	293	32.6
Headings (H)	198	22.0
Formattings (F)	182	20.3
Miscellaneous (M)	0	0

2.1 Data Analysis

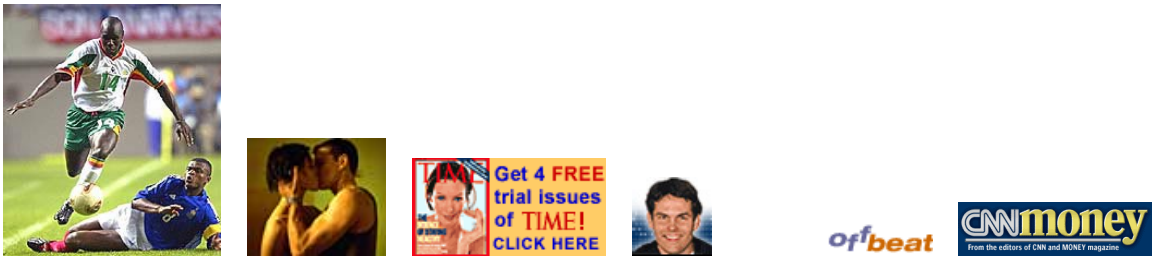
The 25 news front pages contained a total of 899 images. These images were manually annotated using the categories described above. Table 2 shows the distribution of the images per category.

Since our classification categories are functional, the analysis yielded some unexpected results. Three of the numbers in Table 2 stand out in particular: 32.6% for Icons and Logos, 22.0% for Headings, and 12.2% for Commercials. First, the number of advertisements is lower than expected, specially when compared to the number of icons/logos and headings. In addition, while the numbers for the story and preview categories are along expected lines, the number for the formatting category is higher than expectation. The large number of images in the Icons and Logos, Headings and Formattings categories is likely a reflection of the serious efforts devoted by web content providers to maintaining the unique style and appearance of their web front pages.

2.2 Inter-Annotator Agreement

The ultimate goal of this work is automatic categorization of images on the web. However, before machines can be trained to perform this classification, it is important to know whether humans achieve a high rate of consistency on the same task. Therefore, the authors decided to undertake a study on inter-annotator agreement.

Ten out of the twenty-five web news front pages (containing 350 images) were annotate by both authors. Initially, there was disagreement on 67 of the 350 classifications (19.1% error rate). However, after common errors such as those due to lapses in concentration were eliminated, the number of disagreements shrank to 25 (7.1% error rate). In other words, 7.1% of the images were truly ambiguous. An analysis of the errors showed that 22 of the 25 dis-



(a) (b) (c) (d) (e) (f)

Figure 1: Examples of images found on the web. (a) Story; (b) Preview; (c) Commercial; (d) Host; (e) Heading; (f) Icon/Logo.

(a) (b) (c) (d) (e) (f)

Figure 2: Examples of images found on the web with their contexts. (a) Story; (b) Preview; (c) Commercial; (d) Host; (e) Heading; (f) Icon/Logo.

(a) (b)

Figure 3: Sample web images: (a) an image that can only be correctly classified as commercial through text analysis; (b) an image where the associated text is not very informative and image based analysis is needed to achieve correct categorization.

agreements were between the commercial and icon categories with the remainder between the story/preview and host categories.

3. AUTOMATIC CATEGORIZATION OF WEB IMAGES

As seen from the preliminary study described in the previous section, images contained in web publications fall into a diverse set of functional categories. As each category has a different functional role, each demands a different treatment in various web document analysis tasks. For example, to summarize and reformat a web page for display on a small PDA screen, story and preview images (categories S and P) should have the highest priority for transmission, either as is or as down-sampled images. On the other hand, commercial images (C) should have the lowest priority for transmission. For icons and logos (I), and headings (H), it may be desirable to recognize their content (through image recognition for category I and OCR for category H) such that they can be replaced by text labels to save both bandwidth and screen space.

The task of automatic classification of images into these categories ranges from straightforward to highly challenging, depending on the category. Category F (Formattings) can conceivably be identified with high accuracy using simple image features such as uniformity, size and aspect ratio. Many banner ads (a subset of category C) also have distinct aspect ratio. However, the accurate classification into the rest of the categories are much more challenging and requires a combination of sophisticated image understanding and text analysis techniques. Figure 3 shows two examples that demonstrate such challenges. The image in Figure 3(a) is a photograph which could well pass as a story image, however the surrounding text reveals that it is actually an ad. Figure 3(b) shows the reverse situation: here the surrounding text is not informative, only by identifying the image as an graphics image containing mainly text, and further recognizing the text can the image be correctly classified as category H (Headings).

Since the simultaneous automatic classification of images into all of the categories defined above is a very difficult task, a more practical approach is to start with the identification of a subset of these categories, which may have useful applications by itself. Taking this approach, we developed an algorithm designed to specifically identify web images belonging to the S (Story) and P (Preview) categories. As mentioned above, automatic identification of these images would enable a web summarization/re-authoring system to give them higher priority for transmission to a mobile device. It would also be useful for summarization: a news story associated with an image identified as category S is likely to be an important story which should be included in the summary.

A quick study of the image categories reveals that the seven defined categories (we are not counting category M as it is not really defined) can be grouped into two *super classes* . The fist super class, denoted SPA, includes categories S (Story), P (Preview) and A (Host), and the second one, denoted CIHF, contains the rest of the categories: C (Commercial), I (Icons and Logos), H (Heading) and F (Formmatings). The first super class is more likely to contain photographic images of "regular" aspect ratios, and they are often associated with some story. On the other hand, images in the second super class are more likely to be graphic, have "irregular" aspect ratios (e.g, extremely long or wide), and are often not associated with a story.

Based on this observation, we designed our classification procedure as follows. First, a simple size screening process is applied to remove very small images and images of "irregular" aspect ratios. To be more specific, an image is removed if its height is less than 20 pixels or the ratio between the larger dimension and the smaller dimension is greater than 2.5. Our experiments showed that this simple procedure reduces the total number of images to be considered by about half without removing any images from the desired SPA class.

For the remaining images, the main classifier to separate the SPA and CIHF classes is built using both image features and features extracted from the associated text. Then a secondary classifier using only text features is used to further separate out Host images from the SPA class. The remaining images are considered Story and Preview images.

In the following we describe the images features, text features, and the combined classifier in detail.

3.1 Classification of Photographic and Graphic Images

The image characteristic that stands out most at the first glance of a web page is whether an image is photographic (also referred to as "natural"), or graphic (also referred to as "synthetic"). As mentioned above, this feature has strong correlation with the functional categories: an image in categories S, P and A is more likely to be photographic, while the other categories tend to contain more graphic images.

The first question one might ask is whether the format of the image can be used for this classification. The two most common formats used for images in web documents are GIF and JPEG. GIF stands for Graphics Interchange Format and is a lossless format more suitable for graphics. JPEG, developed by the Joint Photographic Experts Group, is a lossy compression scheme more suitable for photographic images. However, despite this distinction the convention of using GIF for graphics and JPEG for photographic images is not always followed for various reasons. In the database described in the previous sections, 14 of the 91 images in category S are photographic images in GIF format, while 13 of the 127 JPEG images are graphic images. Other researchers have also observed similar mixture of image classes within each single format [6]. Thus, image format cannot be used as the primary indication of photographic *vs.* graphic images. It should only be used as secondary evidence to resolve ambiguities after image based classification.

Much of previous research on image based classification in the document analysis community has been focused on the classification of text *vs.* non-text regions within an image [17, 11, 10], predominantly using frequency domain analysis of image intensity.

While this is clearly related to the classification of photographic *vs.* graphic images, there are some important distinctions. On the one hand, the common characteristics used by previous algorithms to identify text regions do not necessarily hold in the broader category of graphic images. On the other hand, a photographic image may very well contain text as part of the scene (sometimes referred to as "scene text").

Some researchers have also explored the color characteristics of different classes of images for classification. Swain *et al.* proposed using features such as degree of color saturation and number of dominant colors to separate photographic and graphic images [6]. Lopresti and Zhou [12] used similar features to identify text regions in web images.

After investigating the two different approaches described above, it became clear to us that these two approaches are complimentary to each other and a combination of the two would likely lead to improved performance. We thus designed a new algorithm incorporating features from both the frequency domain and the color domain.

3.1.1 Frequency Domain Features

Much of the characteristics separating photographic and graphic images are reflected in spatial features of image intensity. For example, graphic images tend to have many sharp edges where as photographic images usually have less well defined regions and exhibit smoother transition between regions. To exploit such characteristics, we derive features from the DCT (Digital Cosine Transform) coefficients of 8×8 subregions (blocks) of an image. Such features have been used successfully before for text vs. non-text image classification [17, 11, 10]. The main innovation in our algorithm is that a clustering procedure is applied first to handle the fact that graphic images in general are much less uniform compared to text images.

The 8×8 DCT results in 64 coefficients. A subset of these are selected using a discriminative analysis carried out on data extracted from a set of training images. First the absolute values of the coefficients are taken, which we will refer to as *absolute coefficients* hereafter. The values corresponding to each absolute coefficient are then normalized by the standard deviation. To estimate the class discriminative power of each coefficient, we compute the within class and between class variances as following. Suppose there are a total of n_p photographic image blocks and n_g graphic image blocks. Let $P = \{p_1, p_2, ..., p_{n_p}\}$ and $G = \{g_1, g_2, ..., g_{n_g}\}$ represent the indexes of photographic and graphics image blocks, respectively. Let α_k refer to the absolute value of the kth DCT coefficient. Let $(\bar{\alpha_k})_P$ and $(\bar{\alpha_k})_G$ represent the means of α_k over set P and G respectively. The within class variance is defined as:

$$\sigma_k{}^2 = \frac{1}{n_p + n_g}\left(\sum_{j \in P}\left((\alpha_k)_j - (\bar{\alpha_k})_P\right)^2 + \sum_{j \in G}\left((\alpha_k)_j - (\bar{\alpha_k})_G\right)^2\right).$$

The between class variance is defined as:

$$\tau_k{}^2 = \frac{1}{n_p + n_g}\left(\sum_{j \in P}\left((\alpha_k)_j - (\bar{\alpha_k})_G\right)^2 + \sum_{j \in G}\left((\alpha_k)_j - (\bar{\alpha_k})_P\right)^2\right).$$

And the discriminative power of the kth coefficient is measured by:

$$\delta_k = \frac{\tau_k}{\sigma_k}.$$

The top $M < N$ coefficients with largest δ_k are then selected as the DCT features.

While DCT features similar to that described above were used directly with success in past efforts to classify an image block as text or non-text, our experiments showed that such a strategy does not work well for photographic and graphic image classification. This is because both categories contain a large range of different image blocks. For example, while graphic images tend to contain sharper edges, they often contain uniform blocks as well. On the other hand, photographic images sometimes contain regions of high frequency variation such as scene text and fences, as well as the more typical smooth-transition areas.

To accommodate the large variation within each class, we apply unsupervised clustering on the training image blocks using the M selected DCT coefficients. To be specific, the K-means clustering algorithm [7] is used to group the training image blocks into a predetermined number of K clusters. Each training image block is then labeled by its cluster index. Finally a normalized cluster histogram is computed for each image, yielding a K dimensional feature. Parameters M and K are chosen empirically and we settled on $M = 18$ and $K = 15$ in our experiments. For classification, each image block is assigned to the cluster with nearest cluster center and the same K dimensional cluster histogram is computed and used as the feature representing the whole image.

3.1.2 Color Features

Swain *et al.* proposed 8 color related features to distinguish graphic and photographic images [6]. A study of those features revealed that many of them are various heuristic ways of implementing aspects of the frequency domain characteristics that are better captured by the frequency domain features described above. We thus selected 2 of the color features that are completely independent from the frequency domain features and thus add most discriminative power. These two features are summarized below for completeness.

- The band different feature: A threshold T between 0 and 255 is selected and a counter C is initialized to 0. For each pixel in the image, if the difference between the largest and the smallest RGB components is greater than T, then the counter C is increased by one. After all pixels in the images have been examined, the band difference feature is calculated as $\frac{C}{S}$ where S is the total number of pixels in the images. This feature has a range of $[0, 1]$ and is a rough measure of the degree of saturation in the image. Graphic images tend to get higher values since they tend to contain purer colors. We chose $T = 50$ as suggested in the original paper.

- The most common colors feature: Given a predetermined number N, the N most common colors in the images are found. Then the feature is simply defined as the fraction of pixels in the images that have one of those colors. This feature again as a range of $[0, 1]$ and is a rough measure of the degree of color concentration. Again, graphic images tend to get higher values since they are often dominated by a few colors. We chose $N = 10$ in our experiments.

3.1.3 Combining the Image Features

There are two possible ways to combine the 18 frequency domain features and 2 color features described above. The most straightforward approach is by directly concatenating the two features, yielding a 20 dimensional feature vector. Our experiments indicated that this approach, not surprisingly, does not perform well. The color features were overwhelmed by the large number of DCT features and did not improve the overall performance.

The second approach, which we adopted, is to use a two stage approach. First a frequency domain classifier is trained using the 18 DCT features. The same classifier is then applied to both training and testing images, giving a classification score for each im-

Table 3: Performance of the photographic/graphic image classification algorithm.

Feature Set	Frequency Domain	Color	Combined
Accuracy (%)	91.1	89.4	92.5

age. This single score is then used as the frequency domain feature, which is concatenated with the 2 color features to form a 3 dimensional image feature. The photographic/graphic image classifier is then trained using these 3 features.

This approach is tested on 462 images collected from the web sites listed in Table 2, which includes 232 photographic images and 230 graphic images. The procedure which produced this dataset is explained in detail in Section 4. The data set is divided into five roughly equal parts, each containing roughly equal numbers of graphic and photographic images. A Support Vector Machine (SVM) classifier [18, 5] was then trained on each four of the five parts and tested on the remaining part. The process is rotated and the combined five part results were then pooled together to arrive at the overall accuracy of the classification algorithm. For SVM classifier training and testing, we used the SVM^{light} system implemented by Thorsten Joachims [8] and tested both the linear kernel and the Radial Basis Function (RBF) kernels.

Our experiments indicated that the RBF kernels performs better than the linear kernel for both the intermediate frequency domain classifier and the final image classifier. Table 3 shows the accuracy achieved using different sets of features. As shown in the table, the combination of the frequency domain features and color features lead to better results than when either feature group is used alone.

3.2 Identifying Story and Preview Images

As described earlier, our strategy is to first build a main classifier to separate the SPA images from the others and then to use a secondary classifier to remove the A images. This sub-section describes the main classifier which uses both image and text features. For the image feature, the classifier uses the output of the SVM classifier used to classify photographic and graphic images (as described in Section 3.1.

A description of the text features follows.

3.2.1 Text Features

Images on the web are almost always accompanied by text and such text often contains useful information about the nature and content of the images. Much research has been carried out in the past on using the associated text for image searching and indexing on the web [6, 15]. For that particular task, it was found that the most relevant text fields are: image file names, image captions and alternate text (defined by the `<alt>` tag in HTML). The functional classification of images is a different problem requiring a different set of features as well as techniques. Since in this case the goal is not to search for a particular image, but rather to classify any given image into one of several broad functional categories, the text fields mentioned above are too specific. Instead, as can be seen from the example images given in Figure 2, the surrounding text of an image (text found in the immediate neighborhood of the image) plays a much more important role in identifying its functionality.

The extraction of the surrounding text of an image is a non-trivial task by itself. Ideally, one should use spatial proximity to judge what text is near a particular image. Unfortunately, while tools for querying spatial information of nodes in an HTML DOM tree are being developed, they are not yet widely available. To get around this problem we used an approximation in our experiments. For each image, we extracted text nodes in the neighborhood of the image node in the DOM tree, within a maximum of 2 levels. A maximum of 20 words each are extracted for "before text" (from text nodes to the left of the image node) and "after text" (from text nodes to the right of the image node). Structural features such as node boundary and whether each node is a hyperlink are preserved during extraction.

For each image, the classifier analyzes the set of extracted text nodes from the neighborhood of the image. The following feature values are computed over the set of text nodes:

Hyperlink Count This is simply a count of the number of nodes that are hyperlinks. Images in class H (Heading) are likely to have larger values for this feature as compared to images in either the S, P, or A classes.

Number Count is a count of the number of all numeric words in the nodes. Images in class C are likely to have larger values for this feature.

Caps Count is a count of the number of capitalized words present in the text nodes. If the first word of a node is capitalized, then it is not included in the count as we assume that it is the beginning of a sentence. Images in the SPA superclass are likely to have higher values for this feature as their contexts usually contain proper names.

Non-dictionary Word Count This feature computes the number of words in the text nodes that do not belong to a dictionary. It is complementary to Caps count feature since most proper names are not found in dictionaries. The dictionary used is WordNet [14], an on-line lexical database developed at Princeton University.

Maximum Words Count This feature computes the maximum number of words in any of the text nodes. Since SPA superclass images are likely to be accompanied by descriptions, the value of this feature for the superclass will likely be high.

3.2.2 Combining Image and Text Features

The image and text features were combined and a Support Vector Machine classifier using SVM^{light} [8] was trained and tested on same training and testing sets used in Section 3.1.3 and described in Section 4. The best results were obtained using the linear kernel of SVM^{light}. The final results are shown in Section 4.

3.3 Identifying the Host Images

Once the SPA superclass is identified, we used a secondary classifier to separate out the host images (A). Since the number of host images in the training and test sets is quite small it was not possible to use a SVM classifier for this secondary task. Therefore, we used a rule-based classifier which consisted of the following two rules:

1. For each text node corresponding to an image, identify the proper names[1] in the node (if any) and then compute the percentage of the words in the node that belong to the person proper names (for example, in "Larry King Live," 66.67% of the words belong to the name "Larry King"). Once this computation is performed for all text nodes corresponding to an image, take the maximum value. If the maximum value is greater than 50%, then proceed to the second rule. Otherwise, the image is not a host image.

[1]We used BBN's IdentiFinder[3] to identify and classify the person proper names

Table 4: Distribution of the images after the initial size screening.

Category	# of Images	Percentage
Story (S)	169	36.6
Preview (P)	45	9.7
Host (A)	26	5.6
Commercial (C)	71	15.4
Icons and Logos (I)	136	29.4
Headings (H)	13	2.8
Formattings (F)	2	0.4
Miscellaneous (M)	0	0

2. For the text node which contains the maximum percentage value (determined in the first rule), check if the node is also a hyperlink. If so, then identify the current image as a host; otherwise, it is not.

The choice of the rules was made with the precision of identifying the host class in mind. We wanted to reduce the number of false positives for the host class as each such instance reduces the recall of the stories and previews class.

Out of 26 host images in the 462 image set, 25 made it past the main classifier. The secondary classifier identified 6 of the remaining host images with 100% precision and 24% recall.

4. EXPERIMENTS

The data collected from the 25 web sites shown in Table 2 consists of 899 images. To increase the training and testing sets, we collected a second set of front pages from the same sites but a different date. The new set consists of 960 images. The resulting set of 1859 images was subjected to the simple size screening test with a threshold aspect ratio of 2.5 (as described in Section 3). After the screening, the resulting set consisted of 462 images. Table 4 shows the distribution of the images in this set.

The set of 462 images was then divided into 5 roughly equal parts containing an approximately equal number of graphic and photo images. Four of these parts are used for training while one is used for testing. The five-fold validation method was employed. In other words, the experiments were run five times where, in each run, one of the five parts was designated as a test set with the remaining four acting as training sets.

The precision and recall numbers achieved by the SVM classifier for the SPA super class are 90.5% and 95.4% respectively. After the rule based host image identification and removal, the final precision and recall numbers for the story and preview images are 82.6% precision and 95.3% recall.

5. CONCLUSION AND FUTURE WORK

As the popularity of the World Wide Web soars, it is increasingly being used as a publication medium that has the ability to instantaneously reach millions of people globally. As a result, web pages now contain an increasing number of images that serve different purposes. In this paper we first described a study of images found in web pages that included defining function based categories for images and statistics on the distribution of images among these categories. We then described an image categorization system that attempts to identify the two most important categories of image: stories and previews. The system uses both image and text features and a hierarchical classification algorithm to achieve precision and recall numbers of 82.6% and 95.3% respectively.

While we focused on the story and preview images for the purposes of this paper, much work remains to be done in function-based image classification of all images. For example, we plan to investigate algorithms to classify images in the heading category so an OCR system can be used to extract the text content. Image recognition techniques need to be explored to interpret images in the icons category. In addition, icons that appear regularly on web sites (for example, newspaper logos) can be classified by analyzing different editions of the pages for repetitions. For the host class, a combination of detecting repetitive images and face recognition will help significantly.

6. CONCLUSION

7. REFERENCES

[1] A. Antonacopoulos and D. Karatzas. An anthropocentric approach to text extraction from www images. In *Proceedings of the Fourth IAPR Workshop on Document Analysis Systems (DAS2000)*, pages 515–526, Rio de Janeiro, Brazil, December 2000.

[2] T. Bickmore, A. Girgensohn, and J. Sullivan. Web page filtering and re-authoring for mobile users. *The Computer Journal*, 42(6):334–346, 1999.

[3] D. Bikel, R. Schwartz, and R. Weischedel. An Algorithm that Learns What's in a Name. *Machine Learning*, 34:1–3, 1999.

[4] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Proceedings of the Tenth International World Wide Web Conference (WWW2001)*, Hong Kong, China, May 2001.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–296, 1995.

[6] C. Frankel, M. Swain, and V. Athitsos. Webseer: an image search engine for the world wide web. *University of Chicago Technical Report TR96-14*, 1996.

[7] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[8] T. Joachims. Making large-scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[9] T. Kanungo and C. Lee. What fraction of images on the web contain text? In *Proceedings of the First International Workshop on Web Document Analysis (WDA2001), online at http://www.csc.liv.ac.uk/ wda2001*, pages 43–46, Seattle, September 2001.

[10] I. Keslassy, I. Keslassy, D. Wang, and B. Girod. Classification of compound images based on transform coefficient likelihood. In *Proceedings of The 2001 International Conference on Image Processing (ICIP2001)*, Thessaloniki, Greece, October 2001.

[11] J. Li and R. Gray. Text and picture segmentation by the distribution analysis of wavelet coefficients. In *Proceeding of The 1998 International Conference on Image Processing (ICIP'98)*, pages 566–570, Chicago, October 1998.

[12] D. Lopresti and J. Zhou. Locating and recognizing text. *Information Retrieval*, 2:177–206, 2000.

[13] N. Milic-Frayling and R. Sommerer. Smartview: flexible viewing of web page contents. In *Poster Collection of The Eleventh International World Wide Web Conference (WWW2002), online at http://www2002.org/CDROM/poster/172*, Hawaii, May 2002.

[14] G. A. Miller. Five Papers on WordNet. Technical Report 43, Cognitive Science Laboratory, Princeton University, July 1993.

[15] E. Munson and Y. Tsymbalenko. To search for images on the web, look at the text, then look at the images. In *Proceedings of the First International Workshop on Web Document Analysis (WDA2001), online at http://www.csc.liv.ac.uk/ wda2001*, Seattle, September 2001.

[16] G. Penn, J. Hu, H. Luo, and R. McDonald. Flexible web document analysis for delivery to narrow-bandwidth devices. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR01)*, pages 1074–1078, Seattle, WA, USA, September 2001.

[17] K. Perlmutter, N. Chaddha, J. Buckheit, R. Gray, and R. Olshen. Text segmentation in mixed-mode images using classificaiton trees and transform tree-structured vector quantization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)*, volume 4, pages 2231–2234, 1996.

[18] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[19] J. Yang, Q. Li, and Y. Zhuang. Octopus: aggressive search of multi-modality data using multifaceted knowledge base. In *Proceedings of The Eleventh International World Wide Web Conference (WWW2002)*, Hawaii, May 2002.